UNIVERSAL
LIBRARY

PRINCIPLES OF THE MATHEMATICAL THEORY OF CORRELATION

PRINCIPLES OF THE MATHEMATICAL THEORY OF CORRELATION

BY

A. A. TSCHUPROW

LATE HONORARY FELLOW OF THE ROYAL STATISTICAL SOCIETY

TRANSLATED BY

M. KANTOROWITSCH, Ph.D., F.S.S.

LONDON EDINBURGH GLASGOW

WILLIAM HODGE AND COMPANY, LIMITED

1939

Printed in Great Britain by Butler & Tanner Ltd., Frome and London

CONTENTS

THE translator wishes to tender his hearty thanks to Dr. J. O. Irwin and Dr. L. Isserlis for the great help they have afforded him in the preparation of the English edition of this work.

AUTHOR'S PREFACE

THE present book is an enlarged reproduction of lectures delivered to the actuarial seminar of the University of Christiania (now Oslo). Although retaining the form of the lectures, the book is divided into chapters complete in themselves, instead of original sections of equal length, necessary for lectures, each of forty-five minutes' duration.

The purpose of this book differs from other works on correlation, inasmuch as its intention is to provide a logical foundation for the theory of correlation and not a guide to the practical application to its methods. Hence the technique of correlation-measurements is hardly touched upon. The development of the fundamental principles of the theory of correlation has not kept pace with the rapid expansion of the methods of its measurement and of their application in all kinds of investigation in the course of the last three decades. This divergence has proved a hindrance both to the theory of correlation and to the advantageous application of statistical investigation in several kinds of scientific inquiry. There are increasingly cogent reasons for the necessity of clarifying the *fundamental notions* and *assumptions* in the Calculus of Correlation ; further, there is a need for clear-cut comprehension of *problems* as well as for close examination of methods of solution applied by the representatives of different Schools.

The present treatise is an attempt to work out *the doctrines of the modern theory of correlation into a homogeneous and comprehensive system from this point of view.*

The presentation is limited to questions which arise from dealing with two variables. Problems arising from the

consideration of three or more variables are for the time being postponed except incidentally.

With this main purpose in view, technical details and statistical and mathematical methods have both had to be set aside as much as possible. I have had to abandon the elaboration of mathematical details in the text. Every competent mathematician will be able to supply the missing steps of the argument without any difficulty, and those not so well mathematically equipped will find in the Appendix the necessary formulae and equations ; the mathematical methods employed are the simplest possible.

By starting with the consideration of the *Discontinuous Distribution* and of the *Laws of Dependence* I have simplified the logical and mathematical ideas needed for the investigation. By such means complications are avoided which would otherwise divert the reader's attention from the essential logic involved by the problems, and it is possible to express the mathematical treatment in terms of *elementary algebra* familiar to the average statistician. The application of differential and integral calculus to the interpretation of ' Normal Correlation ', which it would otherwise be impossible to develop, is unavoidably admitted as an exception.

Contrary to the usual method, which does not make any use of Probability in expounding the Calculus of Correlation, there will be an attempt in the present book to link up the modern theory of correlation organically with the *theory of probability*. Hence this presentation does not start with the numerical reduction of the empirical material but with the analysis of magnitudes with given *a priori* probabilities and their relation to the empirical values observed in random samples. A clear exposition of the rôle of the theory of probability *a priori* in the measurements of correlation is in my opinion the only means of bringing clarity and order into the framework of the theory of correlation. The calculations in which statisticians are engaged achieve their purpose only if there is complete comprehension of what it is they are computing. The actual treatment is

in the main closely associated with the results obtained by the *English School.* Of course, the latter is, so to speak, translated into another mathematical language, and when necessary attuned to the *a priori key.*

I shall not go closely into the logical and philosophical questions which are connected with the notion of Probability. I myself adhere to the School which in the history of the theory is associated with the names '*A. Cournot*' and '*J. von Kries*'. However, I have tried to cast the mathematical presentation into forms which can be filled into other philosophical concepts without very great alteration.

The readers for which a theoretical statistical work is chiefly destined are as varied as the branches of science which nowadays make use of statistical investigation. A well-considered differentiation of the kind of presentation is therefore to be recommended as both the empirical base of statistical research in the fields of social and natural science, and obviously, the resulting methods of practical application are different. As statistical technicalities are left out, the present work is released from the necessity of choosing between the readers of different branches of science. The inquiries involved appear as *preliminary questions common to all those sections of science in which statistical methods are used.* The author's inclination to the social sciences naturally gives an unavoidable bias to the presentation, but he hopes not to be too unfair to the other needs, since he is by his early training not a stranger to natural science.

Finally, one or two remarks as to why the problem of the calculation of the *equation of the correlation surface,* which is the best fit to the empirical data, is not touched upon. The reason is partly because this section of the theory of correlation is still in a rudimentary state and less attention has been paid to statistical research in this direction than to the corresponding problem of the calculation of frequency-curves. Furthermore, the author realizes that

the presentation of these methods can hardly be expressed at present in an easily handled mathematical form. However, the following consideration was decisive: as everybody knows, there is great difference of opinion in the scientific work with regard to the possibility of achieving real knowledge by this method. There are so-called mathematical statisticians for whom the calculation of equations of frequency-curves of correlation-surfaces is the culminating point of the statistical investigation. On the other hand, there are statisticians who consider such calculations a mere pastime and lacking scientific value. If the problem is raised one should not leave out a detailed discussion of the questions, especially as after their elimination the rest would be practically nothing more than the technique of calculation, the description of which is omitted in the present work. Yet the examination of these questions does not belong to the framework of the theory of correlation since they have really no close relation to the measurement of correlation; this could be done more easily and thoroughly by dealing with the problem of determining the equations of frequency-curves, this being the only place where, in the present state of statistical research, the logical analysis can rely on fairly sufficient empirical material.

The list of references which concludes the book does not claim to be a complete Bibliography of the questions considered. Its purpose is rather to refer the reader to those works which are most important for insight into particular problems as well as for guidance to relevant literature.

CHAPTER I

THE MODERN ' MATHEMATICAL ' THEORY OF CORRELATION AND THE METHODS OF ' NON-MATHEMATICIANS '

§ 1

AMONG the functions of statistical inquiry the determination of associations between phenomena under statistical examination plays a prominent part, both because of its importance in various branches of statistics and because of the value it renders to every-day life, so far as the latter depends on numerical expression. From ancient times statisticians have been keenly engaged in the development of those statistical methods which pursue this purpose. In the new era the interest in associations of the particular kind with which statisticians have to deal, has received special stimulus since statistical inquiry has gained ground so rapidly in natural science. The productive impulse of statisticians working in the field of natural science has hereby entered a path which quite significantly deviates from that previously frequented. The fact that students of natural science are more prone to mathematics than those of social science is of great importance. Since the guidance in the struggle for statistical innovation in natural-scientific inquiry lay in the hands of prominent statisticians —Karl Pearson should be mentioned in the first place— the theory of correlation, as this *Novum Organum* of statisticians has become called, has, from its origin, taken mathematical forms which have proved a stumbling-block to advocates of the older conceptions. Thus a most unjustifiable cleavage arose among students of statistics. The so-called ' mathematicians ' sometimes show bias inasmuch

as they disdainfully and without further thought cast aside as lumber the—in their opinion—rudimentary and insufficiently considered ' elementary ' methods of inquiry of non-mathematicians. On the other hand, the ' non-mathematicians ' reject the ' mathematical ' methods of inquiry as being a scientifically sterile ' toying ' with figures, which deludes uncritical minds by a deceptive appearance of precision not in practice attainable, and cannot hold its ground against the criticism of trained statisticians. This unhappy antagonism becomes a serious hindrance to the harmonious development both of the theory of statistics and of those branches of science which apply statistical methods. The termination of such a discussion is certainly to be anticipated. The waves are beginning to grow visibly smoother. As soon as this contentious mood which obstructs mutual understanding is calmed, it will be realized that the greatest cleavage was aggravated by mere mutual exaggeration and that no unbridgeable gap exists between the opinions of the two sides. When this realization comes the bridge can easily be built, and the new ' mathematical ' methods of inquiry will gain prevalence everywhere within their proper limits. Those ' mathematical ' methods of determining associations which at present meet with most passionate resistance, will achieve their object with least trouble because the controversy between mathematicians and non-mathematicians has least real justification at this point, as the new methods are closely connected with the older ' elementary ' methods of inquiry. The modern theory of correlation, pre-eminently indebted to natural scientists, appears on closer examination a logical continuation of ideas fundamentally the same and is rooted historically in the achievements of social statisticians. In order to lay bare these roots it is not necessary to dig too deeply: it is sufficient to cast an unbiased eye over the methods at all times employed by statisticians when inferring associations. For the presentation of the logical construction of the theory of correlation such an adoption of the ' elementary ' methods

of 'non-mathematicians' may be of particular importance for the following reasons: one can overcome the timid distrust of non-mathematicians in the theory of correlation, and furthermore, one can throw light upon the intrinsic property of the tasks which the theory of correlation has to solve by drawing a parallel between these tasks and the old attempts at solution and thus facilitate the comprehension of the logical foundation of the measurements of correlation.

Let us demonstrate by some characteristic examples how strongly the modern theory of correlation is anchored to those methods of inquiry which have been used and praised by the non-mathematicians. This will smooth the ground for the 'mathematical' construction designed by us.

§ 2

In order to enter instantly *in medias res* let us examine more closely the so-called correlation table. Such correlation tables represent in comprehensive form the frequency-distributions containing various combinations of possible values of attributes chosen for the purpose of inferring statistical associations between two phenomena within the field of observation of the inquirer. As a classical example, let us consider a Table showing the age-combinations in which for every age-group it will be recorded in how many cases X-aged men marry Y-aged women. For our formal-methodological consideration let us take as a basis figures derived from an experiment I made. It will not be necessary to enter into details. Under our X and Y must thus be understood any variates with but a single formal limitation that both X and Y can assume only 19 various integral values between 0 and 18.

In which way can we infer from the universe of the observed combinations of the numerical values of X and Y, whether there is an association between them or whether the two variables are independent of each other? Bearing in mind separate values we may observe the most diverse

TABLE 1

		Y																	
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
	0	1	–	1	–	–	–	1	–	–	–	–	–	–	–	–	–	–	–	–
	1	–	–	3	–	–	3	–	1	–	–	–	–	–	–	–	–	–	–	–
	2	–	2	1	1	–	–	1	–	4	1	1	1	–	–	–	–	–	–	–
	3	–	–	–	–	3	2	1	2	–	–	2	1	–	–	–	–	–	–	–
	4	–	–	–	–	1	–	2	–	1	1	1	–	–	–	–	–	–	–	–
	5	–	–	1	1	1	3	1	–	1	1	1	1	2	–	–	–	–	–	–
	6	–	–	1	1	1	–	1	1	–	2	1	–	–	2	–	–	–	–	–
	7	–	2	2	2	–	–	2	–	–	1	–	3	1	–	1	1	–	–	–
	8	–	–	1	1	1	1	3	–	–	–	4	1	1	–	3	–	–	1	–
X	9	–	1	–	–	–	1	3	4	3	2	3	4	4	1	1	1	1	–	–
	10	–	–	2	–	–	1	1	4	4	–	1	2	6	–	2	–	2	–	–
	11	–	–	1	–	–	1	1	1	1	2	–	3	3	2	–	1	–	–	–
	12	–	–	–	1	–	1	1	1	1	2	–	3	1	–	3	–	2	–	1
	13	–	–	–	–	–	1	–	1	1	–	2	–	1	1	–	1	1	1	–
	14	–	–	–	–	–	1	1	2	1	2	1	2	–	–	–	–	1	–	–
	15	–	–	–	–	–	–	–	–	1	1	1	1	–	1	1	–	–	–	–
	16	–	–	–	–	–	–	–	2	–	–	–	1	4	2	–	–	–	–	
	17	–	–	–	–	–	–	–	–	–	–	2	1	–	–	1	–	–	–	–
	18	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1	1	–	1	1

relations between the value X and the corresponding value Y. A larger value X corresponds at one time to a larger and at another time to a smaller value of Y. So on the average one has a clear impression that the Y-values increase with the growth of the X-values; there are a great many exceptions—some quite conspicuous: in one case, for instance, the value 0 of X corresponds to the value 6 of Y, while there is a case where the value 9 of X corresponds to the value 1 of Y. The examination of the results of the separate trials forms no uniform picture. In order to arrive at more cogent inferences one has to work out the results of separate trials suitably—to collate them so that the respective presence or absence of the associations between X and Y should emerge more clearly.

Different statisticians will give different replies to the

question of how to proceed in such work. These divergent answers are substantially justified because, as we shall see, there are several different methods of solving this problem, each one, at times, deserving reference, according to requirements. To a large extent, however, the difference in reply is due to the attitude of the statistician towards 'mathematical' statistics. Mathematical statisticians would suggest certain methods which would be rejected by non-mathematicians. The latter, in their turn, would try to tackle the problem by a method of approach disdained by mathematicians. Let us commence by considering some of those methods which have been devised by non-mathematicians down the years.

A. The fundamental method of non-mathematicians consists in the computation of array-means of one variable —for instance, of the variable Y for increasing values of the other variable, and in determining whether the means increase or decrease with increasing values of X, or whether they fluctuate irregularly round the grand mean of the Y values. In our case we arrive at the following arrays (vide Table 2).

TABLE 2

The value of X . . .	0	1	2	3	4	5	6	7	8	9
The mean value of Y	2·7	4·0	6·25	6·6	7·2	7·1	7·6	7·1	9·1	9·4
The value of X . . .	10	11	12	13	14	15	16	17	18	
The mean value of Y	9·6	9·75	10·8	11·3	9·0	10·8	11·0	11·25	16·0	

We see that the larger is X, the larger on the average is Y—but only on the average, as there are exceptions;

for instance, a noticeable smaller mean value from Y corresponds to the X-value 7 rather than the X-value 6. If we thus apply this method most commonly used by non-mathematicians—in the language of the mathematical theory of correlation it is called ' Calculation of the Empirical Regression Line '—the question which we hoped to avoid in this way is raised afresh, namely, how can one determine whether the mean values of Y fluctuate irregularly or whether they exhibit a tendency towards increase or decrease ? It would be an easy matter to solve the problem if each successive mean value of Y were larger or smaller respectively than the preceding one. The picture is mostly, however, irregular and points—as in our case—to fluctuation in the movement of the mean values of Y. The mathematician, who estimates the equation of the regression line by the Method of Least Squares, has his own, well-developed ways of approach in determining how far it may be considered plausible that the line of regression can have the shape of a straight line parallel to the X-axis. The non-mathematician, who has not such methods at his command, has to seek aid in some other way.

The nearest approach to the foregoing is as follows : to repeat the operation of combining the single values into greater groups until one obtains a uniform picture. If we, for instance, divide our X series into 3 sections (vide Table 2), giving to the first one the values 0 to 5, to the second 6 to 12, and to the last 13 to 18, we arrive at the mean value of Y-values which equals 5·6 corresponding to the mean value of X equal to 2·5 for the first section, for the second one we obtain the mean value of X equal to 9, and that of Y equal to 9·05, and for the last section we get as much as 15·5 for X and 11·6 for Y. Hence with the increase of X-values the Y-values increase without exception : Y is directly associated with X. The association would also appear should we divide our series into 5 sections : we find that the mean value of Y equal to 4·9 corresponds to the mean value of X equal to 1·5 ; the mean value 7·2

of Y corresponding to 5·5 of X; 9·3 of Y to 9 of X; 10·2 of Y to 12·5 of X; and for the highest mean value of $X = 16{\cdot}5$ we obtain the highest value of the Y series $= 12{\cdot}4$. But having further extended the number of groups to 9 one would not observe any unexceptional increase of the values of Y, corresponding to the increase of the values of X.

It is obvious that one can always arrive at a uniform picture by such means. When the series is cut into two, then one mean of values is larger than the other, or both are equal. It is then an easy task to determine whether the association is direct or inverse or whether there is no association at all; but it is likewise clear that no great reliability can be felt in a conclusion derived from division into two sections only, as the suspicion cannot be rejected that mere chance might play its part. The inference of the presence of an association is more conclusive the greater the maximum number of trial groups in which without exception increase or decrease of the one array corresponds to the increase of the other one. The skill of the non-mathematical statistician who selects this method culminates accordingly in the ingenious adjustment of the groups he forms to the wavy regression line, in order to confine successive fluctuations within the narrowest possible bounds. It is not necessary to emphasize the danger of arbitrariness in this connection. Sometimes one tries to guard against this by the arrangement of groups in which certain rules are observed which serve as a standard: for instance, the groups are so constructed that they are comprised of possibly an equal number of single observations, or in such a manner that the scale of values of X is divided into equal portions. Here there is a somewhat primitive attempt to struggle with those difficulties which mathematicians are able to surmount with greater success in a systematically considered way.

B. Another popular method of non-mathematicians is the consideration of deviations of both the arrays from

their respective mean values. Here the starting-point is the consideration of mutual independence, the deviation of X in one direction has to be accompanied by a deviation of Y nearly as often in the same direction as in the opposite one. A large excess of deviations in the same direction would then indicate the existence of a direct association between X and Y, whilst an excess of pairs of deviations of the opposite sign would be evidence of an inverse association. In order to follow this kind of non-mathematical method of inquiry in its systematic development, let us consider an example which is dealt with by Professor G. Jahn in his text-book,* namely, the relationship of the level

TABLE 3

GENERAL DAY'S WAGE. DAY LABOURERS WITHOUT BOARD

Fylker	1910		1915		Ranks			
					1910	1915	Diff.	D^2
Finmark . . .	350	+ 85	+ 89	445	1	1	0	0
Telemark . . .	289	+ 24	+ 34	390	2	4	− 2	4
Troms	286	+ 21	+ 61	417	3	2	+ 1	1
Rogaland . . .	281	+ 16	+ 40	396	4	3	+ 1	1
Vestfold . . .	280	+ 15	+ 18	374	5	6	− 1	1
Vestagder . .	276	+ 11	+ 19	375	6	5	+ 1	1
Nordland . . .	272	+ 7	+ 6	362	7	7	0	0
Buskerud . . .	267	+ 2	− 6	350	8	10	− 2	4
Austagder . .	264	− 1	+ 3	359	9	8½	+ ½	¼
Hedmark . . .	263	− 2	− 9	347	10	11	− 1	1
N. Trøndelag . .	259	− 6	− 17	339	11	13	− 2	4
Hordaland . .	255	− 10	+ 3	359	12	8½	+ 3½	12¼
Østfold . . .	249	− 16	− 21	335	13	14	− 1	1
S. Trøndelag . .	240	− 25	− 40	316	14	17	− 3	9
Akershus . . .	236	− 29	− 39	317	15	15½	− ½	¼
Sogn og Fjordane	233	− 32	− 39	317	16	15½	+ ½	¼
Møre	232	− 33	− 12	344	17	12	+ 5	25
Opland . . .	231	− 34	− 44	312	18	18	0	0
Average wage in Øre	265	—	—	356	—	—	—	—

* G. Jahn, *Statistikkens teknik og metode*, p. 224 (Kristiania, 1920).

of daily wages in different provinces of Norway in the years 1910 and 1915 (vide Table 3). The arrays we are dealing with reproduce the average wage of a male daily wage-earner in the 18 *Fylker* of Norway in both these years, and I arranged the separate districts according to the order of wages in the year 1910. In the first 8 provinces in our Table the daily wages in 1910 were above the general average, and in 1915 only in one of these 8 provinces was the daily wage under the general average for 1915, and in the remaining 7 provinces it was over the average. Out of the last 10 provinces which show under average wages for the year 1910, 2 show a rise above the average for the year 1915, while 8 remain below. Hence out of 18 cases in 15 there may be observed deviations from the average in the same direction; only in 3 cases were the deviations of opposite sign. A direct association is obvious. According to Fechner's * suggestion it may be characterized by an index-number, which is obtained by dividing the difference between the number of deviations in the same direction and those in the opposite one by the sum of these numbers; letting i be this index, in our case we arrive at: $i = \frac{15 - 3}{15 + 3} = + \frac{2}{3} = + 0{\cdot}67$.

The consideration of signs of deviations of values of X and Y from their means does not exhaust what our series of numbers is in a position to reveal in regard to the presence or absence of the association between X and Y. Apart from the direction, the value of the deviation from the mean should also be considered. In the case of independence large and small deviations of X are equally often accompanied by larger and smaller deviations of Y, whether with the same or with opposite signs. If, however, Y stands in direct association with X, then the deviations which are co-ordinated are not only of the same sign but also correspond more or less with regard to their value. On the other hand, if Y stands in inverse association with X, then

* G. Th. Fechner, *Kollektivmasslehre*, pp. 382–5 (1897).

the deviations tend to have opposite signs but are still such that the larger X and Y deviations, on the whole, correspond to one another in absolute magnitude. Hence in the case of independence the algebraic sum of products of deviations taken with their appropriate signs must roughly equal zero ; in direct association it shows a larger positive value, in inverse association a more or less considerable negative value. In our example the sum of the positive products comes to 16,500, and of the negative products equals 45. In this way we also prove that there exists a clear association between the level of wages in the separate provinces in the two years we have considered. We may express this association in summarized form by calculating an index-number in the same way as before, which we shall denote by I : we arrive at $I = \frac{16550 - 45}{16550 + 45} = 0{\cdot}99$.

Now we can proceed one step further. We have just spoken of deviations which correspond to each other in magnitude, and assumed, that in an association, a deviation from the mean in X has a tendency to cause a correspondingly large deviation in Y. This conception may be expressed more precisely : both the variables may differ considerably with regard to their variability. If X shows very great fluctuation, but the values of Y keep within relatively close bounds, then obviously a fairly great deviation in one of the X values from their mean can call forth only a relatively small deviation in the corresponding value of Y ; on the contrary, with small fluctuations of X-values and a considerable variability of Y a relatively small deviation of X-values from their mean produces a considerable divergence on the Y-values. In order to bring both sets of deviations more exactly in relation with each other it is obvious that one must measure the deviations by the corresponding standard deviations. Then their values can really be considered to correspond to each other. Hence let us express the algebraic sum of products of deviations divided by the corresponding standard deviations as r,

then we arrive at an index, which appears in the theory of correlation under the title 'Coefficient of Correlation'. In our example we obtain: $r = 0{\cdot}93$. Consequently the calculation of coefficients of correlation in the same way as the calculation of regression lines can be brought into direct connection with non-mathematical methods. The methods of non-mathematicians contain the seeds of both techniques; yet the full development is only attained in the modern 'mathematical' theory of correlation. Only on this soil is it possible to develop their fundamental conceptions logically and to find firm anchorage in the theory of probability.

C. Finally, let us consider another device which has been adopted in preference by non-mathematicians. Individual values of both the averages are arranged in order so that we allot to the highest value the number (or rank) 1, to the highest but one number 2, &c. The terms are then so arranged that the X-ranks become a successively increasing series. The presence or the absence of an association is here disclosed by the position in which the Y-ranks appear. In a perfectly direct association all the Y-ranks coincide with corresponding X-ranks; in a perfectly inverse association all the Y-ranks form a successively decreasing series. In the case of mutual independence there is no systematic increase or decrease perceptible in the series of Y-ranks. Hence it is necessary to determine each time to which of the three above cases the actual series of Y-ranks approximates. In the association we have just considered between the magnitude of daily wages in different provinces of Norway in the years 1910 and 1915 we see for example that the first and the last ranks occupy the positions expected on the assumption of direct association and that the ranks 7 and 15 are also in their expected positions. This with 18 ranks is evident proof of the presence of a direct association. One could estimate by means of Probability, to what degree it appears improbable that such a coincidence of ranks of both the arrays would emerge in the absence of any association between X and Y. Non-

mathematicians, however, seek to ascertain by some other methods which of the three possibilities—direct association, inverse association, no association—corresponds to the real distribution of Y-ranks. This may be done by recasting the original material as described in A and B above. Still, by setting such problems special methods can be derived. One can, for instance, split up the two series into sections in order to ascertain to what extent the Y-ranks remain within the limits of those sections to which they belong, under the presumption of a direct or inverse association respectively. Accordingly, if we subdivide the 18 provinces of Norway into, say, three equally extensive groups the presence of a direct association emerges quite distinctly (vide Table 3). Although the Y-ranks 1–6 which belong to the first group are not all in the exact positions in which they would stand in a perfect direct association, none of these 6 ranks deserts into any other group ; both the other groups show likewise a picture harmonizing well with the assumption of a direct association, although each group has surrendered one rank, viz. 12 or 13, to another.

The general impression of increasing order in the Y-ranks may be much more clearly and precisely expressed by the use of an index. The differences of corresponding ranks are computed. If the direct association is perfect all the differences are zero. The greater the differences, the less perfect is the direct association. As the algebraic sum of the differences is identically equal to zero, a comprehensive and precise numerical expression must be set up either on the basis of the absolute value of the differences or by a process of squaring. Should we prefer the latter method the sum of square differences would be zero in the case of a perfect direct association. On the other hand, as may easily be shown, in the case of a perfect inverse association the sum of square differences equals $\frac{n(n^2-1)}{3}$, where n is the total number of terms in the series considered. Finally, it is not difficult to prove that when the two series are

completely independent, the sum of square differences reaches the value $\frac{n(n^2-1)}{6}$. Hence, if an index, which we call, according to Karl Pearson, ρ, is defined by the relation $\rho = 1 - [\Sigma D^2] \div \frac{n(n^2-1)}{6}$, then ρ is zero in the case of independence; ρ is equal to $+1$, if there is a perfect direct association; and is equal to -1 if a perfect inverse association is observed. The larger ρ is in absolute value, the more pronounced is the association, whereas the sign determines whether the association is direct or inverse.

In our case, the sum of squared differences equals 65 and $n = 18$; the index is equal to $+0{\cdot}93$. It is accidentally equal to the coefficient of correlation: 'Accidentally', as in general this cannot be expected, the relations between ρ and the coefficient of correlation being rather complicated. They can only be reduced to more manageable formulae under certain assumptions, as K. Pearson succeeded in doing in the case of the so-called 'normal' correlation.

Thus, having been able not only to lay down the general problem of so-called 'rank-correlation', but also to derive the standard Spearman-Pearson formula, we have again reached the domain of the mathematical theory of correlation without having left the ground of methods counted as non-mathematical. A tangible controversy between the mathematical and non-mathematical way of approach has not been noticeable here either. There is no deep gap between them, but the former appears to be the logical sequence and systematic clarification of the latter. Strictly speaking, the elementary methods are also mathematical, as they have to deal likewise with quantities and with quantitative relations. As far as 'Mathematicalness' is concerned there is no deeper contrast between both the methods of inquiry than between the arithmetical relation: $(5 + 3)(5 - 3) = 8 \times 2 = 16 = 25 - 9$, and the algebraic formula: $(a + b)(a - b) = a^2 - b^2$.

§ 3

As we see, the consequent development of non-mathematical methods to determine associations leads us close to the boundaries of the modern mathematical theory of correlation. Moreover, non-mathematicians have prepared the work of the mathematical theory of correlation in a still more essential sense in so far as they have levelled the path for the right conception of the subject of correlation. The original inquiry of non-mathematicians who were primarily interested only in the question as to whether there was an association or not between the phenomena to be investigated, has gradually modified. They began to realize that the series placed before them for examination were distinguishable one from another not only in so far as the association is sometimes clearer, sometimes less distinct, but also inasmuch as the association is sometimes more intense than at other times. Owing to the chance fluctuations of both series the association is always more or less concealed. It emerges with greater clarity when there are fewer chance fluctuations in comparison with the variation of the corresponding terms of both series. Originally the efforts to work out the methods of comparing the series were exclusively devoted to the elimination of the disturbing effect of chance fluctuations. One had nothing else in mind but to bring out the final association as clearly as possible by the reduction of the relative weight of chance fluctuations. We have noticed, moreover, that success is to a great extent dependent upon the kind of association, viz. that there are associations which appear in spite of paucity of observations and the resulting large chance fluctuations, and on the other hand there are those which remain hardly perceptible when the number of observations is large and the chance fluctuations are accordingly reduced. In this way one learnt to interpret the comparison between series in a new sense. The notion of intensity of association as a characteristic and measurable attribute of an association,

as such, began to develop and to be differentiated from the notion of the distinctness with which the association could be detected from the relevant numerical data. This was a decisive step on the path to a rational theory of statistical research along this line. The most important basic ideas of statistical correlation were discovered and thus a firm basis was created for a systematic development of the most appropriate methods. Uncertain trials could now be replaced by a directed and methodical mechanism, the results of which are embodied in the modern mathematical theory of correlation.

CHAPTER II

SUBJECT-MATTER AND PROBLEMS OF STATISTICAL CORRELATION. CAUSAL RELATION AND CORRELATION

§ 1

THE notion of intensity as an objective attribute of associations which are to be statistically determined constitutes one of the foundation-stones of the theory of correlation. It must, however, be logically refined before one can begin to build upon it. This is because at first sight the notion of intensity seems to stand in crass contradiction to the notion of causal relation upon which the deterministic conceptions of our natural science rest and to which most statistical of investigators adhere. The notion of causal relationship includes the assumption that cause and effect are constantly and indissolubly connected: if A is the cause of A', the effect A' follows upon the cause A at all times and everywhere, and never can A' take place unless A has previously occurred. There is no question of a greater or smaller intensity of association: either A is the cause of A' or not—*tertium non datur.* Why is it, then, that we statisticians have exclusively to deal with a more or less intense relationship?

A natural scientist comes across a similar question even when he is not engaged in statistical investigation. The notion of indissoluble relationship seems at the first glance to exclude all quantitative connections between associated phenomena which have not the form of direct proportion. If A and A' are indissolubly connected, then upon A always follows A', upon $A + A$ follows $A' + A'$ and upon nA follows nA'. It seems that relations of another kind are

impossible. Yet in the province of exact natural sciences the greater part of the work of inquiry is devoted to the task of revealing the true form of functional relationships, and forming mathematically precise laws which the phenomena follow. How is this contradiction to be explained?

The explanation is quite simple. If a scientist is engaged in an inquiry in a field not yet thoroughly investigated and where relations have not yet been fixed, he is not always in a position to pick out from the magnitudes he has measured those which show just causes and their proximate effects. Assume, for instance, that without the slightest knowledge of the relations between the weight and mass of bodies one undertakes an empirical investigation of the relations ascertainable by measurements made with a balance. One constructs for the purpose of the investigation a number of different large regular cubic dice of material as homogeneous as possible and one examines their measurements. When weights and volumes have been measured, one discovers that they are directly proportional. On the other hand, if instead of volumes the length of the edge of the dice is measured, which might just as well be done if one has no knowledge of the real relations, one arrives at another mathematical law: the weight, within the range of errors of observation, is proportional to the cube of the measured values. If one has measured surfaces of dice instead of the length of the edges, one would have discovered a third law. One sees from this example that it is sheer good luck if magnitudes chosen blindly for measurement are directly proportional to one another. It is really more probable that one will come across functional relationships of quite a different nature; for one kind of measurement which leads to a direct proportion many others may be found which give quite different results. Of course, the choice is seldom entirely blind. As a rule, one possesses some preliminary knowledge. Yet it is an exception, when one proceeds with certainty. One need not be surprised, therefore, that among the laws laid down by physicists, chemists,

and other students of natural science, that of direct proportion is not the most prevalent.

In the same way the other apparent contradiction may be removed: the appearance of non-indissoluble relationships within the sphere of an inquirer's contemplation. In practical research work one has continually to deal not with non-indissoluble relationships but with more or less loose ones. An undeniable relationship exists between the attributes of parents and their offspring, between the barometer reading and the height above sea-level at which it is measured. Yet if one considers the individual measurements from which these relationships are derived, then one has a very confused picture before one's eyes: sometimes a son of far below average size issues from a giant-like father, sometimes, on the contrary, where the father is a dwarf, the son is a giant; barometrical readings everywhere fluctuate between such wide limits that sometimes a pressure will be recorded at the seaside which falls far short of that at stations situated considerably higher. How can this be reconciled with the assumption of the rule of indissoluble relationships between cause and effect?

One assumes that the cause A is indissolubly connected with the effect A', the cause B with the effect B', &c., so that at all times and everywhere A takes place A' follows, and A' is not found anywhere, unless A has previously taken place. Now, if one contemplates the relatively complicated phenomena X and Y and seeks to determine their relationship, it might happen that X appears as a combination of A and B, and Y of A' and B'; then X and Y likewise appear to be indissolubly connected. But if X is a combination of A and B, whereas Y is a conceptual unit which in addition to A' and B' contains C', then, though one will never be able to observe Y without X having been previously observed, yet one will notice that some other effects follow upon X—for instance, an effect Z, which is conceptual unit comprising A', B', and D'. Conversely, in a case where X includes some surplus components

which are missing in Y, one will remark that Y always follows X ; but also some other cases, apart from X, precede the effect Y. And finally, in a case where X appears to equal $A + B$ and Y to equal $A' + C'$, one is able to observe both how the effects of the cause X differ from Y and the causes of the effect Y differ from X.

Corresponding to this simple scheme are the causal relationships in the examples which one used to illustrate the notion of Mathematical Probability : the tossing of coins, drawing of balls from closed urns, and other so-called games of chance. From a closed urn, containing two balls which are marked with a number, one ball is drawn : the ball No. 1 is just as likely to appear as the ball No. 2. The complex of causes we are considering is not indissolubly connected with the effects we have in view : it has not a single effect but two different possible effects. This can be explained by the fact that we have tried to bring a relatively concrete effect into relation with a complex of causes much simplified in comparison. All are in themselves causally uniquely determined events, and it would be an easy task to segregate causes and effects which stand in indissoluble relation to one another. If we substantially simplify the effect we have in view, then the indissolubility of the relationship is restored : from our urn containing two balls at each drawing a ball is extracted and not a die —the appearance of a ball is an effect indissolubly connected with the complex of causes we are considering. On the other hand, we arrive at indissoluble relationships when we modify the complex of causes approximately : if we state in addition how the balls have been placed in the urn and the movements made by the hand which draws the ball, then the extraction of a specific ball appears as an effect of the complex of causes thus formed and the extraction of the other ball will no longer belong to the possible effects of the cause we are considering. The possibility of producing the indissoluble relationship exists in all such cases. Yet often our interest is not centred on such indissoluble

but insignificant relationships, but just on the relationship of chosen parts of the entire complex of causes with their concrete effects, regardless of the fact that such a relationship is not indissoluble.

Of course, the above-mentioned scheme is not applicable to all causes of non-indissoluble relationships. I only wished to demonstrate by this simple example how the occurrence of such non-indissoluble relationships can be brought into conformity with the assumption that cause and effect are always indissolubly connected with each other. When working with experimental material in a sphere not yet thoroughly investigated, one operates with hypotheses which have to be developed and interpreted by the inquiry undertaken; one is then in a position to select the phenomena which one is trying to connect—our X and Y—so that, though they may contain elements which are, causally, indissolubly connected, they are not comprised exclusively of such. Then one must expect a relationship which is no longer indissoluble but more or less intense.

§ 2

The greater or lesser intensity of relationship between X and Y can partly be accounted for by their composition: the greater the weight of the causally related elements the more intense is the relationship. If, say, $X = A + B + C$ and $Y = A' + B' + D'$, the relationship is then more intense than in the case where $X = A + B + C$ and $Y = A' + D' + E'$. The hereditary relationship between father and son is, for instance, more intense than that between grandfather and grandson. The length of the left arm and that of the right one of the same individual stand in a more intense relationship than the length of the arms of two brothers; again, the relationship between the corresponding attributes of brothers is, in turn, more than that of cousins.

However, the intensity of relationship is still not determined by the composition of X and Y. With the same

composition, the relationship may be more or less intense according to the variability within the inquirer's field of observation of the elements of X and Y which are not causally related. The relationship between volume and weight appears perfect when bodies of the same homogeneous material are compared. If, however, bodies of different material are drawn, then the relationship is more or less loose according to circumstances. A small stone ball may weigh more than a considerably larger wooden ball; the greater the difference in density of the bodies to be measured the looser is the relationship between volume and weight. If exclusively stone balls or exclusively wooden balls are measured, one expects a more intense relationship than when some of the balls measured are of stone and some of wood. Consequently a relationship between two non-indissolubly connected phenomena may sometimes be more, sometimes less intense. It may even be the case that the actual composition within the inquirer's field of observation of the elements of X and Y which are not causally related conceals the causal relationship between X and Y, or apparently in the case a relationship, really non-existent. If, say, some stone balls and some larger wooden balls are accidentally accessible for the purpose of inquiry, then one is unable to disclose any direct relationship between volume and weight. If balls of heavy dark wood and equally large balls of light coloured aspen-wood are compared, then an influence of colouring on the weight may be disclosed which actually does not exist. These imaginary examples may be paralleled by countless instances from actual practice. The statistics of Russian compulsory fire-insurance disclose a striking relationship between the average number of buildings destroyed in one conflagration in the country and the use or non-use of fire-engines for its extinction: fires extinguished by a fire brigade furnished with a fire-engine are, on the average, more destructive than others. To conclude from this that the destruction of fire-engines constitutes the best means of reducing damage from fire would be as absurd

as to suggest that, in order to diminish their weight, all yellow objects should be painted white—from the observation that gold, yellow-coloured balls weigh more than silver, white ones. The simple explanation is that only the larger villages have fire-engines, and they are seldom used for small fires concerning only a few houses.

These simpler examples are sufficient to show that the problems for the investigator who intends to inquire into relationships between phenomena which interest him are becoming more abundant in scientific practice than one would have supposed if one had paid exclusive attention to indissoluble relationships and concluded that there is no relationship between X and Y if they are not indissolubly connected with each other. It is insufficient to ascertain whether X and Y are associated with each other or not. If a relationship is traceable it must be elucidated more closely: the law of relationship must be determined as precisely as possible and the intensity of relationship must be appropriately represented; but first of all the kind of relationship must be closely explored, particularly the influence of elements not causally related. The interpretation of a relationship is often the most important part of inquiry. Results, achieved by formally identical treatment of sets of data which superficially appear exactly the same may be fundamentally different in their meaning, and if one does not pay due attention to this fact, one may come to recommend the destruction of fire-engines in order to diminish the ravages of fire. Suppose, for instance, that a barometric reading is made first at various heights above sea-level and secondly at various distances from the seashore. The series of numbers do not in themselves reveal in each case the kind of distinction existing between the first and second measurements. The inquirer who works out the numerical data will deduce from the sets of figures exactly corresponding information regarding the law of relationship. But the content and logical value of what he discovers in this manner are quite different in each case:

in the case of relationship of barometric reading with the height above sea-level the point in question is the law of decline of atmospheric pressure, and in the other case this particular method merely brings into relief the environs of the shore where measurements have been made. If one had taken another direction when walking inland from the coast one would have come upon an entirely different law.

The importance of the right interpretation of observed relationships cannot be too strongly emphasized. Within the field of statistical research which always deals with a complex jumble of causally related and causelessly coincident phenomena, the inquirer must not stop after having ascertained that there is a relationship between his X and Y: he must use his utmost endeavour to fathom what the observed relationship really means and what is his real basis. In cases in which results achieved by statisticians are utilized for practical advice and decisions, the half-knowledge of relationships which remains without any elucidation or which are even wrongly interpreted, is often worse than ignorance. In confinements where a physician assists as obstetrician the ratio of still-born children is, on the average, higher than in those where child-bed is attended by a midwife with no medical assistance. Nevertheless, the husband who, in a difficult case, renounced medical assistance on account of this statistical relationship would be bringing trouble on his own head. Russian statistics of fires show a relationship between the fluctuations of the number of buildings burnt down yearly and the harvests of these years. Damage caused by fire increases in years of bad harvest. The relationship is very distant. Yet what does that mean? The inquirer who first discovered it was of opinion that there was an immediate influence of the harvest upon the frequency of fires. If this is the case, by taking measures to improve agricultural technique among peasants, one is using the best means to alleviate the immense burden laid on Russian Economy by enormous damages from fire. Actually there are pre-

sumably other reasons, viz. the relationship of harvest, on the one hand, with damage by fire and, on the other with the atmospheric conditions of the year. In those districts which are responsible for a poor Russian harvest, dry years are the years of lean harvest and drought encourages fires. Accordingly, one would expect a more plentiful harvest from the improvement of agricultural technique, but hardly less damage from fire. Extensive investigations, undertaken by the Central Statistical Board of Soviet Russia under the management of N. Tschetwerikoff to explain the influence of meteorological factors upon the harvest in various Russian districts, have disclosed, *inter alia*, a remarkable relationship between the harvest of winter cereals and the rainfall during the last few weeks before its sowing. The relationship may be traced as much to the influence of rain upon the quality of the soil, as to the damaging effect of rainy weather upon the seeds which will later be used for sowing. Should the second explanation be confirmed by further statistical and experimental inquiries in agriculture, one will then be able to improve the harvest by taking care to obtain better seed material; but if the former interpretation is correct, quite other measurements might be appropriate.

If one asks how the problems to be dealt with can be solved, in the case of non-indissoluble relationships, it is clear, in the first place, that scholastic inductive methods cannot be applied to such relationships. For the latter rest on the supposition that the relationship between X and Y, should it exist at all, must always be indissoluble; they infer that a definite effect Y is related to a definite cause X by eliminating all other phenomena which might be considered possible causes of Y; this is done by identifying cases where either Y exists but these other phenomena are missing, or, conversely, one of these other phenomena are in existence but Y fails to appear. This conclusion looses every justification as soon as one drops the supposition that Y must be independent of X in case Y is not indissolubly

connected with X. A different method of inquiry must be applied to determine such loose relationships.

It is clear, likewise, that the methods to be applied cannot be based on a negative attribute of the relationships to be ascertained, viz. that they are not indissoluble. Positive attributes of these non-indissoluble relationships must be used. The property which non-indissoluble relationships possess of being more or less intense, is particularly suitable as a basis. Therefore the measurement of intensity of relationship becomes a central problem in the theory of methods which have in view the comprehension of non-indissoluble relationships. Under the comparatively simple conditions of the above scheme of formation of non-indissoluble relationships, mathematical probability proves to be a suitable measure of intensity of relationship. In more complicated cases the notion of mathematical probability is not sufficient to make possible a measurement of the intensity of a relationship. Another scale more suited to the nature of the problem has to be sought.

Greater or smaller intensity is a specially conspicuous attribute for non-indissoluble relationships. The inquiry can, however, be concerned with finer features of them. In this way a fertile, ordered system of methods arises, the rational development of which forms the subject of the theory of correlation. In order to gain a systematic view of the whole of these methods as they appear at the present stage of knowledge, we must start from a more exact version of the idea of non-indissoluble relationships. We must transform this idea in connexion with the idea of mathematical probability into the quantitatively precise notion of the ' stochastic ' * connexion between chance variables, which forms the essential basis of all our developments.

* I use ' stochastic ' (for the Greek verb στοχάζεσθαι, to presume) as a synonym of ' based on the theory of probability ' (Wahrscheinlichkeitstheorie). Vide J. Bernoulli, *Ars Conjectandi*, p. 213 (Basileac, 1713), and L. von Bortkiewicz, *Die Iterationen*, p. 3 (Berlin, 1917).

CHAPTER III

STOCHASTIC CONNEXION AND FUNCTIONAL RELATIONSHIP BETWEEN VARIABLE MAGNITUDES

§ 1

In order to derive clearly the fundamental notions of the theory of correlation it is proper, in the first instance, to disregard all concrete subordinate features and to consider the problem in abstract mathematical form. To build a sure foundation we must start from exactly formulated definitions.

A magnitude which can assume with definite probabilities different values we will call '*a chance variable of the* k*th order*'. The set of its possible values and of their respective probabilities we shall call the '*frequency distribution of chance variables*'. In dice-throwing, value of the figure turned up is, for example, a chance variable of the 6th order, as it may, with equal probability of $\frac{1}{6}$ each, assume the values 1, 2, 3, 4, 5, and 6.

The notion of a chance variable is a particular case (*genus proximum*) of the general mathematical notion of a variable magnitude, whereas the existence of the frequency distribution appears as a specific difference (*differentia specifica*). A single-digit figure is a discontinuous variable which can take 10 different values from 0 to 9. It becomes a chance variable of the 10th order in the case where it takes these different values with definite probabilities. Its frequency distribution may take various forms according to the design of experiment. Suppose the arrangement makes all digits equally probable, then the frequency distribution is expressed by the values 0, 1, &c., up to 9, and by the probabilities

of those values, all of which are equal to $\frac{1}{10}$. The number of white balls in a set of 20 balls is a variable which may assume 21 different values from 0 to 20. It will become a chance variable of the 21st order when we add that the 20 balls are drawn from a closed urn with just as many white balls as non-white balls, every ball drawn being replaced before the next extraction takes place: this is because under such circumstances, corresponding to each number of white balls out of the 20 balls drawn, there is a definite probability easily calculated according to the well-known rules of Probability. It will become a chance variable of the 21st order, with another frequency distribution, if the number of white balls in the urn is not half the total number of balls but one-quarter or two-thirds, as also in the case where the balls drawn are not replaced.

This example of drawings from an urn is well suited to explain not only the notion of the chance variable but also its importance in scientific research work. The inquirer frequently has to obtain his material in a manner corresponding to his drawing from a closed urn. In Vital and Social Statistics, for instance, we have sometimes to deal with so-called random sampling which exactly resembles the scheme of drawing from the urn with or without the replacement of the balls drawn. The plankton investigator carries from the ocean bed small samples of the tiny fauna which populates it in order to infer from these samples the contents of its immensurably extensive ' urn '. The physician draws a drop of blood from the patient's body, dilutes it, and then under the microscope counts the blood corpuscles in a very small fraction of the diluted solution, in order to become acquainted with the properties of his patient's blood necessary for his diagnosis. The number of red and white corpuscles in the field of a haemacytometer have the properties of a chance variable of exactly the same kind as the white and non-white balls which we considered in the examples above.

The application of the sampling method to natural science

is not confined to cases where the investigator uses sampling intentionally. On closer examination one sees that the inquirer has often to deal with samples, even where this was by no means intended. When the botanist discovers a new flower and counts the petals of the specimen he has picked he is, strictly speaking, in the same position as if he had drawn and read the figures on a ticket from an urn containing a number of tickets marked with figures, some different, some the same. If he gathers another specimen of his newly discovered flower, he will perhaps find the same number of petals but more likely another number. The same flower may exhibit different petal numbers in different specimens, odd or even, differing from one another by one or more units. Professor C. V. L. Charlier, for instance, has counted the petals of 321 specimens of *Trientalis Europaea* from the neighbourhood of Lund* : the greater part—nearly half—had 6 petals ; but over one-third of the specimens counted had 5 petals, over one-twelfth 7 petals, and in two samples he found as many as 9 petals. In lilac, which has a considerable preponderance of 4-petalled blossoms, one occasionally comes across blossoms with either 5 or 3 petals. Hence even in the case where the botanist gathers hundreds of specimens he is still, so to speak, drawing tickets from the urn of nature ; he has many samples —but still only samples before him.

We can go still further. Suppose the botanist succeeds in procuring all the specimens of his newly-discovered flower available in the world at any one time, will not his specially exhaustive inquiry still be a sampling inquiry in another sense, namely, in regard to the generative renewal from year to year ? In all such cases the investigator has to deal with samples of samples, just as the physician, when counting blood corpuscles under a microscope, is considering samples of the drop of blood drawn as a sample from the human body.

* Vide E. Czuber, *Die statistischen Forschungsmethoden*, pp. 115–116.

This consideration leads us to a deeper conception of the importance of the notion of chance variables in scientific inquiry. Chance variables can appear within an investigator's field of observation not only as a means to an end —as a result of a method of working selected after deliberate forethought—but also as a directly preassigned object of inquiry which, as such, lies within the environment we must investigate. The number of petals of *Trientalis Europaea* is a chance variable which most frequently assumes the value 6, but it may also have other values between 5 and 9, whereas the probabilities of values exceeding 6 rapidly decrease with the increase of the values. The stature of an adult Norwegian is a chance variable which takes all possible values within pretty wide limits with probabilities which decrease symmetrically on both sides of the most frequent stature with an astounding regularity. To go further into the question of causal mechanism upon which the occurrence of such chance variables is based would divert us too far from our proper object. For our purpose it is sufficient to prove that chance variables, the notion of which we have exactly defined, are not playthings idly constructed by mathematicians, but that they really do dog the footsteps of the scientific research worker, sometimes as a resource, sometimes as a proper object of inquiry.

The twofold significance of the notion of chance variables for the inquiry, namely, the circumstance that the chance variable can arise as a means to an end as well as being an end in itself, is of the greatest importance to the theory of statistics. Let us consider some other examples which throw light upon this distinction from another angle.

Measurement, loaded with errors of observation, is one instance in which the inquirer becomes aware of the chance variable. Suppose, having forgotten our Euclid, that we consider the question of the sum of the angles of a triangle as a problem in natural science and wish to solve it experimentally by measurement. For a great number of triangles all three angles are measured and for each triangle the sum

of the three is computed. The true values of the sum is, as we know, 180 degrees in each case. Yet the values of the individual sums differ sometimes more and sometimes less from 180 degrees. The 'measured' sum of angles appears to us not as a constant but as a magnitude which, when all measurements are carried out in the same way with the necessary care, takes different values with definite probabilities; it is a chance variable in the sense of our definition.

Another example. Somebody wishes to ascertain his height on his twenty-first birthday, and with the help of a friend takes careful measurements. What interests him is not a chance variable but a quite definite magnitude: his height expressed in centimetres, millimetres, &c., on the selected day. But the measurement will not give him exactly his true height; because of unavoidable errors of measurement the result of the measurement will deviate sometimes more and sometimes less, sometimes to one side, sometimes to the other, of the required magnitude. The measured stature will appear to be chance variable with frequency-distribution determined by the technique and skill of the measurer. In order to arrive at the required true height one is forced to consider this chance variable and to treat it appropriately. But in such circumstances it is not an object of inquiry. It is necessary to know it and its frequency-distribution, only to be able to determine the numerical value of the true height as exactly as possible and to estimate as reliably as possible the accuracy of the determination.

On the other hand, let us suppose that the object of our interest is not the stature of an individual of 21 years of age on his birthday but the stature of 21-year-old Norwegians and that for the purpose of its determination the statures of a number of Norwegians of 21 years of age are measured. In this case not only the result of the measurement but also the magnitude to be measured is a chance variable. The true statures of Norwegians to be measured

are different and their various values follow a definite law of distribution—the so-called Gauss-Laplace's law of errors —so that the stature of a Norwegian of 21 years of age appears as a chance variable in the real sense of our definition of the notion. And the measured stature is, in its turn, likewise a chance variable, its frequency-distribution being determined by the technique and the skill of the measurements, and at the same time by the frequency-distribution of the values of the true body-height. One of these chance variables—the measured stature—is also in this case but the means to an end, but the end is now a different one, namely, the knowledge of the true stature and its frequency-distribution. We must now adapt our method of inquiry to our altered purpose.

§ 2

1. The idea of stochastic connexion of variables which is fundamental to the statistical theory of correlation and must be precisely distinguished from the notion of functional relationship, depends on the notion of the chance variable. If the variable Y is functionally related to the variable X, after the determination of the value of X there is no further room left for chance in the determination of the value Y. If Y equals X^2 and X equals 4, then the value of Y is 16; if Y equals the square root of X and X equals 4, then although the value of Y may equal $+2$ or -2, neither of these values has a definite probability, and the conclusion as to whether Y should be put equal to $+2$ or -2 cannot be decided by chance, but depends on considerations. On the other hand, if after the determination of the value of X, Y appears as a chance variable capable of taking various values with definite probabilities, then we are faced with a ' stochastic connexion ' between Y and X. For instance, if Y denotes the sum of the numbers thrown with one white and one red die, and X the number thrown with the white die, then X and Y are stochastically connected, because for each given value of X, Y may take six

different values with the same probabilities, according to the result of the throw of the red die ; if X equals 3, then Y may equal 4, 5, 6, 7, 8, or 9 ; if X equals 5, then Y can equal 6, 7, 8, 9, 10, or 11.

If X is a chance variable and Y is functionally related to X, then Y likewise is a chance variable, but only as long as Y is considered without any reference to the value of X. For instance, if X is a figure turned up by the throw of a die and $Y = X^2$, then X can take six different values—1, 4, 9, 16, 25, and 36, each with a probability 1/6. But for each given value of X, Y is no longer a variable, but has a perfectly definite value ; if X equals 3, then Y equals 9 ; if X equals 5, then Y equals 25.

The notion of stochastic connexion can be generalized to apply to the case of any number of variable. If when the values of X, Y, Z, and U are fixed the value of the variable T is uniquely determined, or can take several different values which have no definite probabilities, then T is functionally related to X, Y, Z, and U. On the other hand, T is stochastically connected with X, Y, Z, and U when, after the fixing of the values of X, Y, Z, and of U, T can take different values with definite probabilities. Let us, for instance, denote by T the sum of the numbers thrown with three different-coloured dice by X the number thrown with the white die by Y that thrown with the red. If we assume that X and Y have definite values, T can take six different values, each with a probability 1/6 as 1, 2, 3, 4, 5, or 6 is the figure shown by the throw of the third die.

It is most important for the statistician to note the possibility that T may be stochastically connected with X considered separately or with Y considered separately, but when the values of both the variables X and Y are fixed it loses the property of a chance variable. If T denotes the sum of the numbers thrown with one white and one red die, T is stochastically connected with the number thrown with each individual die. As soon, however, as

both numbers are fixed, T is also uniquely determined: if the white die turns up 1 and the red one 6, then T takes the value 7 and can no longer take any other value; the relationship has become a functional one.

Mutual connexions of this kind between three variables form a contrast to the functional relationship of one variable with two independent but not chance variables. Let us consider, say, the relationship of the boiling-point of a solution of common salt in water to the concentration of the solution and the pressure at the surface of the water. In this case the combination of the values X and Y uniquely determines the value of T: at a given concentration and a given pressure the boiling-point is fixed. But there is no tangible relationship between boiling-point and pressure when concentration remains indefinite. There is no reasonable answer to the question—at what temperature does a solution of common salt begin to boil under atmospheric pressure? The question remains meaningless as long as no closer determination of concentration is available. This determination may be precise, as we have supposed above. We then have a functional relationship between boiling-point and the other two determining factors. Yet it is quite possible that this closer determination is made in such a way as to make concentration appear a chance variable with a definite frequency-distribution. If one asks what is the boiling-point under atmospheric pressure of a solution of common salt taken at random from the store-room of a laboratory, then the question is scientifically valueless but not at all meaningless, provided that the store-room comprises a number of vessels containing solutions of common salt of different degrees of concentration. The answer then will be: the boiling-point is a chance variable with the following precisely specified frequency-distribution. And such a question does not appear scientifically valueless in all cases. One has only to remember the random drawing of samples from Nature's store-room. One comes across chance variables of such origin much

more frequently than would appear to be the case at first sight: they cannot always be easily recognized as such. For example, chemists unhesitatingly state the atomic weight of a lead exactly as they state the atomic weight of any chemical element. But we now know that lead with this atomic weight is a mixture of substances with different atomic weights which are combined in the earth's crust, but are obtainable separately in laboratories. And so the atomic weight of lead recorded in Chemists' Tables is actually a chance variable of the same kind as the boiling-point of a salt solution taken at random from the store of a laboratory. As, however, the different ingredients of the mixture we term lead occur in nature in only slightly fluctuating proportions, the atomic weight of lead, with our limited precision of measurement, can hardly be distinguished from a constant. Recent progress in natural science has given us many examples of such mean values of chance variables masquerading as constants.

The clear distinction between the ideas of 'stochastic connexion' and 'functional relationship' is the first step towards understanding the theory of statistical correlation as distinct from the study of natural law. The inquiry of natural law sometimes takes the form of statistical investigation, but the presentation of functional relationship as precisely as possible is the purpose it pursues, whereas the task of statistical correlation is always to ascertain the characteristic features of the stochastic connexion between variables. We shall have many opportunities of going in detail into the consequences which follow from the two different aims.

2. Stochastically connected variables can arise in the same way as separate chance variables. They can likewise appear as a means to an end as well as an end in themselves. Non-chance variables which are in functional relationship with each other can be transformed into stochastically connected chance variables when their measurement is affected by errors of observation. The boiling-point of pure water and

the pressure at the water's surface are functionally related to each other: a definite boiling-point corresponds to a given pressure, a definite pressure to a given boiling-point. However, there is no functional relationship between boiling-points measured by a physicist and the corresponding pressures, but they are stochastically connected with one another: if the measurements are sufficiently numerous one will be able to ascertain that the same measured pressure sometimes appears coupled with a higher and sometimes with a lower boiling-point, and that the same boiling-point sometimes corresponds to a higher and sometimes to a lower measured pressure.

In a case where boiling-point and pressure are both non-chance variables standing in functional relationship to each other, they will appear after measurement as chance variables. It may also occur that the measurement may transform one variable only into a chance one, and that the true values of the other may be known. Suppose, once again, that we have entirely forgotten our Euclid and that we wish to discover by experiment the formula which connects the sum of the angles of a polygon with the number of its sides. The angles of a number of polygons—triangles, quadrilaterals, pentagons, &c., are measured and the results of the measurements for each polygon are summed; the number of sides is ascertained by counting, which, with care, gives perfectly precise values. Now, if one considers the values of the sums of the angles for a definite number of sides, one will find for a polygon of n angles instead of $180\ n - 360$ degrees, sometimes larger, sometimes smaller values which when measurements are carefully carried out prove to follow a definite frequency-distribution. Here to the true values of one variable correspond a set of values —taken with definite probabilities—of the other. This is an essential difference from the relationship between boiling-point and pressure. Both kinds of inquiry have in common that what occupies the investigator's mind is a functional relationship between non-chance variables.

Suppose, on the other hand, that we are considering an investigation into the relationship between the body-height and chest circumference of Norwegians of 21 years of age or into the relationship between the stature of fathers and their sons. Here not only the results of the measurements but also the true magnitudes to be measured are stochastically connected chance variables ; to a given stature there will correspond different values of chest circumference not only in the tables of our measurements but also among the individuals measured. There are those of tall stature with narrow chests as well as those with short legs and a well-developed thorax ; sons, though descendants of one father, have not all the same stature. Even if we considered the true values of the magnitudes concerned, we would still find no functional relationship.

Hence we see that the investigation of stochastically connected variables involves a considerable variety of problems. Sometimes the stochastic connexion is only a veil behind which functional relationship sought is hidden ; sometimes it is just the stochastic connexion which the inquiry must elucidate. The methods of inquiry must also be adapted to the end pursued. If we are to find our way to the law of functional relationship between magnitudes which interests us by considering stochastically connected variables, then it will be through methods which really belong to the sphere of theory which has been described systematically although not exhaustively in the Theory of Errors of Observation. On the contrary, the statistical theory of correlation has to develop scientifically all possible methods of elucidating the stochastic connexion between the relevant magnitudes. We shall now consider closely what we understand by the scientific elucidation of stochastic connexion. Then we shall return once again to the contrast between the methods of the theory of correlation and those which we have just assigned to the theory of adjustment.

§ 3

If we are asked the purpose of investigating of a chance variable or several stochastically connected chance variables, the answer is to be found in the nature of the objects of investigation.

All that can be said about a chance variable is available when its frequency-distribution is determined. Everything else is deducible from the frequency-distribution. To ascertain all the possible values of the variables and their corresponding probabilities is accordingly the real task. In this form, however, our knowledge of chance variables is indeed complete but not easily manageable nor sufficiently lucid. It must be properly condensed to be applicable to our purpose. It is difficult to compare two frequency-distributions directly without simplifying their properties. By means of aptly constructed and comprehensive coefficients all that is worth knowing can be obtained in workable form from the frequency-distribution.

Of these comprehensive coefficients the first to be considered are the mathematical expectation and the standard deviation of chance variables. By *mathematical expectation* one understands the mean value of all possible values of a variable weighted with their respective probabilities: if the chance variable X can assume the different values X_1, X_2, . . ., X_k with the probabilities p_1, p_2, . . ., p_k then the mathematical expectation of X is defined as $\sum_{i=1}^{k} p_i X_i$: we shall denote it by EX. The *standard deviation* of X is defined as $\sqrt{\sum_i p_i [X_i - EX]^2}$; the usual symbol for it is σ_X. The square of the standard deviation is called '*variance*'.*

* *The Translator's Note:* Verbally: 'The square of the standard deviation—I call *Streuung*.' This term has been introduced by Tschuprow and is the obvious equivalent of R. A. Fisher's term 'variance'.

The value of the mathematical expectation of a chance variable fixes the mean position round which individual values of the variable cluster at a greater or smaller distance. The amount of scatter of individual values round this mean position is indicated by the standard deviation or the variance. A variance different from zero is an essential feature of a chance variable: the variance can vanish only when all possible values of the variable are equal, i.e. when it is not a chance variable but a constant.

In order to consider only a few examples to which we will often return, let us suppose that X indicates a number thrown with a die. The mathematical expectation of X in this case equals $\frac{1}{6}[1 + 2 + 3 + 4 + 5 + 6 = 3{\cdot}5]$; the variance of X is $\frac{1}{6}[\frac{25}{4} + \frac{9}{4} + \frac{1}{4} + \frac{1}{4} + \frac{9}{4} + \frac{25}{4}] = \frac{35}{12}$. If Z denotes the sum of the numbers thrown with dice, we obtain, similarly, $EZ = 7$ and $\sigma_Z^2 = \frac{35}{6}$.

It is not necessary for the time being to consider coefficients which indicate comprehensively the asymmetry and other fine features of the frequency-distribution.

§ 4

1. The investigation of two or more stochastically connected variables proceeds in the same way. The intrinsic property of the stochastic connexion between two chance variables consists in the appearance of the possible values of one variable in combination with different possible values of the other variable, with a definite probability for each such combination. We shall denote the set of different combinations of possible values of both variables and their corresponding probabilities by the '*joint frequency-distribution of the variables*'. The notion of the joint frequency-distribution is easily extended to the case of any number of stochastically connected chance variables.

If the joint frequency-distribution is given, then one knows all that can be stated about the stochastic connexion between variables. All the rest can be deduced from the joint frequency-distribution. Accordingly the determina-

tion of the joint frequency-distribution may be considered to be as the real object of the inquiry. For the same reason that certain characteristic comprehensive coefficients are usually considered in place of, or rather, as a supplement to the joint frequency-distribution, in the investigation of a single chance variable; similar coefficients, which are completely characteristic of the joint frequency-distribution, play a prominent part in the investigation of several stochastically connected chance variables.

2. In order to survey easily and systematically various methods which can be applied to the investigation of joint frequency-distributions, we must become acquainted with a number of ideas bound up with the notion of stochastic connexion.

The set of values which the variable Y can take when the variable X has taken one of its possible values, and their corresponding probabilities, I call '*the conditional joint frequency-distribution*' of values of Y for the given value of X. The mathematical expectation of Y, the standard deviation of Y, &c., calculated from the conditional joint frequency-distribution, are called '*conditional mathematical expectation*', the '*conditional standard deviation*', &c. If the conditional mathematical expectation of Y is expressed as a function of the corresponding value of X, then the resulting analytical expression is called the '*regression equation*' of Y upon X. In graphical terms we speak of regression lines. If the conditional standard deviation or the conditional variance of Y is expressed as a function of the corresponding value of X, we adopt the terminology introduced by K. Pearson and speak of '*scedastic equations*' and '*scedastic lines*', respectively, from the Greek verb—*σκεδάννυμι*, I scatter. If the conditional standard deviation of Y remains constant for all values of X, then the association of Y with X is called '*homoscedastic*'; otherwise one speaks of '*heteroscedasticity*'.

Let us consider the notions '*regression*' and '*scedasticity*' more closely in the light of an example. Let X

be the number thrown with a white die, Y that with a red die and T the sum of X and Y. Further let $E^{(i)}T$ be the conditional mathematical expectation of T which corresponds to the value X_i of X and $E^{(j)}X$, the conditional mathematical expectation of X, which corresponds to the value of T_j of T. The value of $E^{(i)}T$ is composed of the relevant values of X, i.e. X_i, and the value of the mathematical expectation of Y, i.e. 3·5. The regression equation of T upon X is, accordingly, $E^{(i)}T = 3{\cdot}5 + X_i$; the regression of T upon X is consequently linear. The regression equation of X upon T is likewise easily calculable. The possible values of T go from a minimum of 2 to a maximum of 12. T can only be equal to 2 if each of the dice turns up 1; hence the variable X in this case can take only one value, and its conditional mathematical expectation must equal this value; it thus becomes equal to half the corresponding value of T. T can only be equal to 3 if the white die turns up 1 and the red die 2, or vice versa; the two combinations are equally probable; the conditional mathematical expectation of X thus comes to $\frac{1}{2}.1 + \frac{1}{2}.2 = 1{\cdot}5$ and equals half the corresponding value of T. In the same way we can satisfy ourselves that the conditional mathematical expectation of X equals half the corresponding value of T in the remaining cases. Hence the regression equation is $E^{(j)}X = \frac{1}{2}T_j$; consequently the regression of X upon T is linear, likewise.

The conditional standard deviation of T which corresponds to the value X_i of X is determined by the fluctuations of Y and, at all values of X, is equal to the constant $\sigma_T = \sqrt{\frac{35}{12}}$; consequently T is homoscedastically associated with X. With regard to the conditional standard deviation of X, where T equals 2 or 12, it is equal to zero. When T is equal to 3, X with the same probabilities can take the values 1 and 2; the deviations from the conditional mathematical expectation of X, which equals 1·5, come to $\pm \frac{1}{2}$, the squares of deviations $\frac{1}{4}$ each. The variance equals $\frac{1}{4}$,

the standard deviation equals $\frac{1}{2}$. Evidently the standard deviation has the same value when T is equal to 11, since X can only take the values 5 and 6, with equal probabilities. In the same way the conditional variance for $T = 4$ can be shown to be $\frac{2}{3}$, &c. Hence the association of X with T is heteroscedastic.

3. If the conditional frequency-distribution of Y remains the same for all possible values of X, I call Y ' stochastically independent ' of X. It can then be shown (cf. Chap. IV, § 1) that X must likewise be stochastically independent of Y in the sense of the conditional frequency-distribution if X remains the same for all possible values of Y: consequently two chance variables cannot be other than mutually independent. If the conditional frequency-distribution of Y changes in any way as the variable X goes through all its possible values, then Y and X are not stochastically independent.

The notion of stochastic independence is one of the foundation-stones of the theory of statistical correlation. The first step in the investigation of stochastically connected variables is to inquire whether they are mutually independent. If, in the sense of the above definition, they are, then it is unnecessary to worry further: the frequency-distribution is then determined by the frequency-distributions of the separate variables. On the other hand, if the variables are not independent of one another, then the frequency-distribution must be studied more closely and represented in a proper way.

The definition of independence with which the theory of correlation has to work can also be formulated in another way. Several different formulations have been suggested. In certain circumstances some of them may render valuable service provided that the notion concerned is exactly definite and distinguished from other competing definitions of independence. Further, it is desirable to introduce different technical terms for the different definitions of independence. The introduction of definitions which pursue special pur-

poses is justifiable when thus restricted. The above-mentioned definition of the idea of stochastic independence is, however, the best basis for the general theory of statistical correlation. It is the most stringent of all formal definitions of independence. If, on this definition, two chance variables are independent of one another, then they are also independent on any differently other formal definition of the notion. It must be emphasized here that we are speaking of mathematical definitions. It is very important to bear in mind that independence with which the theory of correlation is concerned is a technical term and is related in quite a special way to independence in the usual causal sense. Suppose, for instance, there are six closed urns filled with white and black balls, bearing the numbers 1, 2, 3, 4, 5, and 6. By throwing the die we will determine from which urn draws shall be made : if a 3 turns up, then the balls are drawn from urn No. 3, &c. The number of balls to be drawn from the urn—n—remains constant. The proportion of white balls drawn which can take $n + 1$ different values, namely $0, \frac{1}{n}, \frac{2}{n}$, &c., up to 1, and the number thrown are two stochastically connected chance variables. If the proportion of white balls is different in the different urns, then the two variables are not independent. If all the proportions are equal, then the two variables are mutually independent in the sense of our definitions. The causal mechanism which connects the events is much the same in both cases. In interpreting the results of the measurement of correlation, one must always pay the very greatest attention to this special definition of the idea of stochastic independence with which the theory of correlation operates. Otherwise one comes to conclusions which fall far outside the scope of statistical correlation and one becomes involved in contradictions which compromise statistical methods.

Of the other definitions of independence which are of value to the theory of correlation, I will only mention K.

Pearson's original definition of ' uncorrelated '. K. Pearson calls the variable Y correlated with X when the conditional mathematical expectation of Y takes different values for different values of X ; if, however, the conditional mathematical expectation of Y remains constant at all values of X, then Y is said to be uncorrelated with X. Hence the non-correlation of the variable Y with the variable X is expressed by the fact that the regression line of Y on X, which graphically represents the conditional mathematical expectation of Y as a function of X, lies parallel to the X-axis ; if the regression line of Y upon X is not a straight line parallel to X-axis, then Y is correlated with X.

This definition of ' uncorrelated ' is of considerable value to the theory of correlation and we shall often have to make use of it. It must be carefully distinguished from stochastic independence as defined above. If the variable Y is stochastically independent of X, then it cannot be correlated with X in the sense of K. Pearson's definition. But if Y is uncorrelated with X it does not follow that Y is stochastically independent of X ; the conditional mathematical expectation of Y can remain constant for all values of X, but the conditional frequency-distribution of X can change in some other manner, when the variable X goes through all of its possible values ; the conditional standard deviation can, for instance, have different values for different values of X. Suppose that the above-mentioned six urns contain the some proportion of white balls, but that the number of draws to be made from the urn chosen by the throw of a die is proportional to the figure turned up ; that it is to be n when the die turns up 1, $2n$ when it turns up 2, &c. Under these conditions the proportion of white balls drawn is uncorrelated with the number thrown, for the mathematical expectation of the proportion is determined by the relative number of white balls in the urn concerned, which is the same for all six urns. But the conditional standard deviation does not remain constant ; when the die turns up 4 it is half as

great as when the die turns up 1. Hence the proportion of white balls is not independent of the number thrown.

It must further be remembered that in contrast to stochastic independence, which is always mutual for both variables, K. Pearson's 'uncorrelated' implies no mutual relationship between the variables: from the fact that Y is uncorrelated with X it must not be inferred that X is uncorrelated with Y. The regression of Y on X can take the form of a straight line parallel to the X-axis, and the regression of X on Y can nevertheless take a form deviating arbitrarily from the straight line parallel to the Y-axis. In order to bring out this important fact quite clearly let us consider an example.

From a closed urn containing an equal number of white and black balls, k series of draws are made, replacing the ball after each draw. The number of draws is determined by chance, perhaps by drawing a ticket from another urn containing a series of numbered tickets, the figure on the ticket being used to fix the number of draws to be made, after which the ticket is replaced so that the probability of all possible values of the number of balls drawn remains constant. By n_i we denote the number of draws in the ith series and by w_i the proportion of white balls drawn. Let us examine both the chance variables—n and w—in relation to their association, under the supposition that the numbers of the tickets fluctuate within fairly wide limits so that the numbers of draws in the individual series are sometimes quite small and sometimes very great. If, in the first place, one considers the relation between the magnitudes w and n, one sees immediately that under the given assumptions the conditional mathematical expectation of w for each value of the number of draws in the series remains equal to $\frac{1}{2}$: w is not correlated with n and the regression of w on n is represented by a straight line which is parallel to the axis of n. On the other hand, if one considers the conditional frequency-distributions of n-values which correspond to the different values of w, one obtains a regression

curve of quite a different character ; to very small values of w as well as to values greatly in excess of 0·5 there corresponds a small number of draws in the series in question ; on the contrary, to values of w which differ by a small amount from 0·5 there corresponds a larger number of draws. The regression of n on w is thus represented by a curve with an ascending and descending branch. In the well-known text-book by G. U. Yule* we find the lines of regression of w on n and of n on w illustrated by a concrete example—namely, by a correlation between the sex-ratio of the newly-born in various registration districts of England and Wales and the number of births in the districts in question. The picture corresponds exactly to the above-mentioned scheme.

§ 5

We have now arrived at a point which is of fundamental importance in understanding the nature of the difference between statistical correlation and natural law.

The regression equation of Y on X expresses the functional relationship between the conditional mathematical expectation of Y and X. It corresponds, formally, to the natural law which expresses the functional relationship between two variables. Sometimes the regression equation may have the same meaning as the natural law. Let us again consider the experimental determination of the relationship between the sum of the angles of a polygon and the number of its sides. Suppose that the errors of observation follow the Gauss-Laplace's law of errors. The sum of the measured angles of triangles, of quadrilaterals, &c., are chance variables and their mathematical expectations are the ' true ' values in question : 180 degrees of triangles,

* *The Translator's Note :* The author refers to the G. Udny Yule, *An Introduction to the Theory of Statistics*, p. 176, edition 1911. In the new edition (11th, by G. Udny Yule and M. G. Kendall ; London : Charles Griffin & Co., 1937), the curve is placed on p. 213 (Fig. 11.10) and the corresponding table (11.6) on p. 202.

360 degrees of quadrilaterals, &c. The functional relationship between the mathematical expectation of the sum of angles which corresponds to a given number of sides, and the number of sides coincides in this case with the inquired law : the sum of the angles is equal to 180 degrees multiplied by the number of sides less two.

Yet the law of nature is always reversible. If Y is an explicit function of X one can express X as an explicit function of Y by means of formal-mathematical operations and suitable symbols : if $Y = X^2$, then X equals the square root of Y ; if $Y = aX + b$, then $X = \frac{1}{a}Y - \frac{b}{a}$. On the contrary, the regression equation of Y on X and the regression equation of X on Y are not deducible from each other. The regression of the sum of the numbers thrown with two dice on the number thrown with one die is expressible, as we have seen, by a linear equation $E^{(i)}T = 3{\cdot}5 + X$. The regression equation of the number thrown with one die on the sum is likewise linear, but it has the form $E^{(j)}X = \frac{1}{2}T_j$. By no ingenuity of mathematical reasoning can one equation be deduced from the other : each must be obtained independently by the consideration of the joint frequency-distribution. This is itself by no means surprising, since the regression equations do not connect the same magnitudes : the one connects the conditional mathematical expectation of Y with X, the other, the conditional mathematical expectation of X with Y ; they have just as little in common as an equation which connects X and Y with another equation which connects two variables, U and W. However, the inquirer who has turned from Natural Science to Statistics has in mind the functional relationship between Y and X, to which the regression equations seem to refer, and this peculiar and irreversible relationship is a stumbling-block to him. He can but with difficulty overcome the impression that this is an inherent imperfection of one of the usual ways of treating stochastically associated variables which must be removed by the calculation of a unique

regression equation which functionally connects X and Y and which permits one to express Y as a function of X as well as X as a function of Y. Such efforts provide evidence of a misunderstanding of the nature of stochastic connexion. They must not, however, be rejected without further consideration as, within certain limits, they are not ill founded. They must, however, be kept within these limits.

We can determine where these lines may be fixed by considering the double part which may be played by consideration of stochastically associated variables. Where the stochastic connexion appears as a shell hiding its kernel—the functional relationships between the true magnitudes which concern the investigator—the latter rightly feels unsatisfied by obtaining regression equations. This gives him no definite results to his inquiry. What he is anxious to know is the true functional relationship between the true values of X and Y; the conditional mathematical expectations of X and Y in themselves do not interest him at all; the formulae which express their functional relationship to corresponding values of other variables are of value to him only as auxiliary to his efforts to determine the law which connects the values of X and Y. If, for instance, he wishes to discover the law which connects the boiling-point of water with the pressure, it will be of no use to him to set the regression equations which connect the conditional mathematical expectation of the boiling-point with the pressure and the conditional mathematical expectation of the pressure with the boiling-point; they do not express what he is anxious to know. His problem remains unsolved so long as he is confined to regression equations. Only when, by means of suitable treatment of the stochastically associated variables before him, he succeeds in advancing to the functional relationship concealed behind the stochastic connexion can he consider he has achieved his end. The manner in which this should be done is a separate question, the consideration of which is of recent date, but into which it is unnecessary to enter here. It is sufficient

for us to have realized that the effort to obtain a unique equation representing the law of functional relationship between X and Y is in such cases justifiable and that such an equation is different in nature from equation of regression.

The position is quite different when stochastically connected chance variables are the real object of the inquiry. If the magnitudes X and Y under investigation are chance variables with an intrinsically stochastic connexion between them, then there is no functional relationship between X and Y at all: definite values of one variable cannot be brought into correspondence with definite values of the other variable; if X has a certain value, then Y can accept a number of different values with definite probabilities and none of these values has more right than any other to be considered the value corresponding to the value of X. In this case, unlike the previous one, it is not only difficult to discover an equation determining Y from X, but it is unnecessary to search for it at all, as it is non-existent. It is possible to present the mutual association between X and Y in different forms: they are exhaustively described by the frequency-distribution; their relevant features are comprehensively summarized by regression equations and in other ways, which we will consider later; but an equation presenting Y as a function of X, or X as a function of Y is not among them: this kind of representation is different in nature from the notion of stochastic connexion.

Now we can sum up. There are cases where the conditional mathematical expectation of Y coincides with the true value which is functionally related to X and corresponds to the given value of X; under such circumstances the regression equation of Y on X gives directly the required law of functional relationship between Y and X. There are also cases where neither the conditional mathematical expectations of Y nor those of X coincide with the true values of the magnitudes under investigation which are functionally related; then the required law of functional

relationship can be represented neither by the regression equation of Y on X, nor by the regression equation of X on Y, and can only be obtained by methods suitable to the particular problem. Finally, there are cases where there can be no question at all of a functional relationship between the magnitudes under investigation, when their mutual relations are such that they cannot be represented by a ' law ' in this way. Here regression equations express definitely and as well as possible certain relevant properties of the connexion between the magnitudes under investigation. One must always bear in mind the multiplicity of tasks imposed upon the inquirer when considering stochastically associated variables if he is to make a conscious and rational choice of the method to apply, and above all to arrive at a correct interpretation of the results. A trained critical sense in this respect is one of the most important conditions of success. The real object of the investigation must be continually borne in mind.

The theory of statistical correlation has, then, to deal with cases where the stochastic connexion of two or more chance variables is under investigation. The statistical methods of inquiry aim at as complete a comprehension as possible of mutual associations between the stochastically connected chance variables, and at as practical a representation as possible of their most important characteristic features. Now we have become more closely acquainted with the particular nature of the task, we can venture to take a further step from this secure position into regions not otherwise without danger, and first of all survey systematically all the methods which provide a comprehensive representation of stochastic connexion.

CHAPTER IV

THE *A PRIORI* JOINT FREQUENCY-DISTRIBUTION AND THE RELATED SYSTEM OF PARAMETERS AND COEFFICIENTS

§ 1

STATISTICAL correlation deals with stochastically associated chance variables. We have seen that the particular methods which aim at grasping and developing the idea of stochastical connexion are the result of the conceptional distinction between stochastic association and the functional relationship more familiar to the student of natural science. We shall now survey the main features of these methods systematically, paying special attention to the case of two stochastically associated chance variables.

If we ask in what way it can most clearly be brought to light, whether the variables are mutually independent or not, and in the latter case by which methods their connexion can be best comprehensively characterized, the answer must be cast into mathematical form. Although we need not carry out complicated calculation, we cannot dispense with algebraic formulae: without their support all formulations would become either too long or too involved or they would remain too hazy.

Suppose we have two chance variables, X and Y. We denote by $p_{i|}$ the probability that the variable X assumes a particular one of its k possible values—namely, the value X_i — ; by $p_{|j}$ — the probability that the variable Y takes a particular one of its l possible values—viz. the value Y_j; by p_{ij} — the probability that Y takes the value X and Y the value Y_j simultaneously. Let us further denote by $p^{(i)}_{|j}$ the conditional probability that the variable Y takes

the value Y_j when the variable X has taken the value X_i; and by $p_{i|}^{(j)}$ the probability that the variable X takes the value X_i when the variable Y has taken the value Y_j. As the probability of occurrence of two events which are not independent is equal to the product of the probability of the one by the conditional probability of the other, we have the relations

$$p_{i|j} = p_{i|}p_{|j}^{(i)} \qquad p_{i|j} = p_{|j}p_{i|}^{(j)}.$$

According to our definition (cf. Chap. III, § 4, 3) the variable Y is independent of X in a case where the conditional frequency-distribution of Y remains the same for all values of X, i.e. if the conditional probability of any value of Y for every value of X is equal to the unconditional probability of the same value of Y, or put into the language of formulae if

$$p_{|j}^{(i)} = p_{|j} \text{ for } i = 1, 2, \ldots, k \text{ and } j = 1, 2, \ldots, l.$$

Conversely if $p_{|j}^{(i)} = p_{|j}$ for every possible value of i, and for every possible value of j the conditional frequency-distribution of Y remains the same for all values of X and the variable Y is independent of X.

If in the relation $p_{i|}p_{|j}^{(i)} = p_{|j}p_{i|}^{(j)}$, true in all circumstances, we put $p_{|j}^{(i)} = p_{|j}$, we obtain $p_{i|}p_{|j} = p_{|j}p_{i|}^{(j)}$, and hence $p_{i|}^{(j)} = p_{i|}$. If thus $p_{|j}^{(i)} = p_{|j}$ for all values of i and of j, then is also $p_{i|}^{(j)} = p_{i|}$ for all values of i and j. The independence of the variables X and Y results from the independence of the variables Y and X, as was pointed out above without any proof (cf. Chap. III, § 4, 3).

In a similar way we convince ourselves that in the case of the mutual independence of variables, $p_{i|j} = p_{i|}p_{|j}$ for all possible values of i and j.

Conversely, if $p_{i|j} = p_{i|}p_{|j}$ for all possible values of i and j, then $p_{|j}^{(i)} = p_{|j}$ at all possible values of i and of j and the variables X and Y are mutually independent.

§ 2

1. In the case of mutual independence of the variables all differences $p_{i|j} - p_{i|}p_{|j}$ are equal to zero. If one or more

differences diverge from zero the variables are not independent. The greater the difference the more the relation between the variables deviates from the independence. Accordingly, the magnitude of the differences expresses an essential property of the connexion under consideration. It forms, therefore, the foundation of one of the main groups of methods which serve to represent stochastic relations.

When both X and Y can take only two different values each, all four differences are equal in absolute magnitude. If we put $p_{1|1} - p_{1|}p_{|1} = \delta$, we have identically

$$\delta = p_{2|2} - p_{2|}p_{|2} = -[p_{1|2} - p_{1|}p_{|2}] = -[p_{2|1} - p_{2|}p_{|1}].$$

The value of δ appears in this case as a convenient numerical characteristic of the relation between the variables. If we insert in the expression for δ

$$p_{1|} = p_{1|1} + p_{1|2},\ p_{|1} = p_{1|1} + p_{2|1}$$

and note that $p_{1|1} + p_{1|2} + p_{2|1} + p_{2|2} = 1$, then we further obtain $\delta = p_{1|1}p_{2|2} - p_{1|2}p_{2|1}$. In this symmetrical shape δ is particularly readily used for the formation of coefficients.

When the variables may take more than two values the differences $p_{i|j} - p_{i|}p_{|j}$ need not be all equal. A single one among them cannot in this case serve as a criterion of the relation. Comprehensive coefficients must be based on the utilization of all the differences. Since the sum of the differences is identically equal to zero, the next thing to be done is to proceed from the squares of differences. On this foundation various coefficients can be constructed. The greatest importance is to be attached to the coefficient introduced by Karl Pearson, which he calls ' Mean Square Contingency '. We shall denote it by φ^2 and define it as

$$\varphi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{[p_{i|j} - p_{i|}p_{|j}]^2}{p_{i|}p_{|j}}.$$

If the variables are mutually independent, all differences $p_{i|j} - p_{i|}p_{|j}$ are equal to zero and consequently also $\varphi^2 = 0$. Consequently φ^2 can be equal to zero only if all differences are equal to zero ; from $\varphi^2 = 0$ it follows directly that the

variables are mutually independent. If the relationship between X and Y is a unique functional one, to each value of X there corresponds with it a definite value of Y; if the value of Y which corresponds to the value X_i of X is denoted by Y_i, then the probability $p_{i|j}$ is equal to zero when j is different from i; as regards the probability $p_{i|i}$ it is equal to $p_{i|}$. φ^2 therefore assumes the value $k - 1$, where k denotes the number of possible values of the variable.

In the value of $\frac{1}{\sqrt{(k-1)(l-1)}}\varphi^2$ we have accordingly a coefficient which is equal to zero in the case of mutual independence of the variables—and only in this case!—and assumes the value 1 if the variables stand in any form of unique functional relationship to one another. The numerical value of $\frac{1}{\sqrt{(k-1)(l-1)}}\varphi^2$ keeps within the limits 0 to 1 and is nearer to the upper limit the more the form of the conditional frequency-distributions of Y approach the form in which the probability of one of the possible values of Y reaches the level of certainty and the sum of the probabilities of the remaining values becomes vanishingly small, or, in other words, the less marked the chance character of the variables after the determination of the value of X, the closer are the mutual associations between Y and X to the type of unique functional relationship. Thus in the magnitude $\frac{1}{\sqrt{(k-1)(l-1)}}\varphi^2$ we possess a measure for a property of stochastical connexion between Y and X, which is highly relevant to the inquirer.

In the case where X as well as Y can take only two different values each, $\frac{1}{\sqrt{(k-1)(l-1)}}\varphi^2$ is reducible through simple transformations to the form

$$\frac{1}{\sqrt{(k-1)(l-1)}}\varphi^2 = \varphi^2 = \frac{\delta^2}{p_{1|}p_{2|}p_{|1}p_{|2}}$$

(vide below, § 7).

2. All coefficients which are constructed from the values of δ and φ^2 have a characteristic feature in common ; they utilize only the probabilities of the possible values of X and Y as well as of their different combinations, and ignore these values themselves. Whether the possible values of X and Y are great or small, whether they fluctuate within a wider or narrower range, has no influence on the value of the δ and on the value of φ^2, provided only the probabilities $p_{i|}$, $p_{|j}$, and $p_{i|j}$ remain the same. This makes the group of coefficients suitable for the examination of cases where one may speak only with certain reservations of stochastically associated chance variables in the sense of our definition. Not wanting the possible values of X and Y for the calculation of φ^2, we need not know them—and not needing to know them we need not measure them ; indeed, it is, at bottom, irrelevant whether they are measurable at all or not : it is a matter of no consequence that they are expressed in numbers ; they must be distinguished from each other only in so far that we count the various combinations. We may even go a step further and assume that the different categories of both variables are distinguished not quantitatively but qualitatively. The possibility of calculating the value of mean square contingency remains unaltered by this. If, for instance, we make an inquiry in marriage statistics with regard to the association between the religion of bridegroom and that of bride, we are able to calculate the mean square contingency according to the above-mentioned formula exactly as in the case where we investigate the association between the age of bridegroom and that of bride. Religion is certainly not a chance variable magnitude in the sense of our definition (cf. Chap. III, § 1). It cannot be considered as a variable magnitude at all, whether as a chance or non-chance one. Yet it is a variable qualitative characteristic. And among variable qualitative characteristics we can distinguish chance variables from non-chance variables as well as among variable magnitudes : where definite probabilities appertain

to the different non-measurable gradations corresponding to the qualitatively different categories of a characteristic, we can denote the characteristic concerned as a *chance variable characteristic*. In the same sense the notion of stochastical association must also be widened by extending it from chance variable magnitudes to chance variable characteristics, which can be either measurable or non-measurable, quantitative or qualitative. Two *variable characteristics* are *stochastically associated* with each other when, fixing the category of one characteristic, the other remains a chance variable in the sense that it is capable of falling into one of several categories with definite probabilities. On the other hand, if it loses the character of a chance variable characteristic after the particular category of the first characteristic has been fixed, then the two characteristics are not stochastically associated with each other, but are related in a manner which corresponds to the functional relationship between variable magnitudes.

As an example of a non-quantitative chance characteristic, the colour of balls to be drawn from a closed urn may be cited. If one further assumes that the balls differ from one another not only in colour but by any other distinguishing marks, then one is faced by an example of stochastically associated non-quantitative chance characteristics.

By using the generalized notions of the chance variable characteristic and the stochastical association of chance variable characteristics, we can express more precisely the special nature, already mentioned, of methods which proceed from the consideration of the differences $p_{i|j} - p_{i|}p_{|j}$ — namely, that they are suitable for the examination of stochastical association, not only between chance variable magnitudes but also between chance non-measurable and even non-quantitative characteristics. This is an essential advantage of this group of methods. Other methods which we will now consider are only adoptable without any further consideration to the investigation of non-quantitative

characteristics if the two chance variable characteristics have only two categories each (cf. below, § 7). Otherwise their application to non-quantitative characteristics presupposes an artificial ' quantification ' of them : one tries to arrange the different categories of the qualitative characteristics in such a way that the succession of terms of the series may with some justification be interpreted as an increase or decrease of an underlined quantitative characteristic ; it is of some importance that the conceptually quantitative gradations do not permit of too arbitrary an estimation.

§ 3

1. Another main group of methods which are used in the examination of stochastically associated variables begins with the consideration of conditional frequency-distributions. If variables are stochastically independent of each other, then the conditional frequency-distribution of Y remains the same at all values of X. The missing independence appears through various characteristic numbers which characterize the conditional frequency-distributions of Y, undergoing a definite change when the variable X runs through the series of its possible values. The more precise formulation of the way in which various characteristic numbers of the conditional frequency-distribution changes thus forms one of the characteristic features of the law of dependence of X and Y.

In order to survey these methods methodically we must introduce a number of new symbols. Let us denote by $m_{f|g}$ the mathematical expectation of the product of the fth power of X by the gth power of Y, so that

$$m_{f|g} = Ex^f y^g = \sum_i \sum_j p_{i|j} x_i^f y_j^g.$$

Let us further introduce the symbols

$$\mu_{f|g} = E[x - m_{1|0}]^f [y - m_{0|1}]^g =$$

$$= \sum_i \sum_j p_{i|j} [x_i - m_{1|0}]^f [y_j - m_{0|1}]^g \text{ and } r_{f|g} = \frac{\mu_{f|g}}{\mu_{2|0}^{\frac{1}{2}f} \mu_{0|2}^{\frac{1}{2}g}}.$$

Putting $g = 0$, we obtain the appropriate parameters by which the frequency-distribution of X is characterized. If we put $f = 0$ we obtain parameters of the frequency-distribution of Y. Thus $m_{1|0}$ is the mathematical expectation of X, $m_{0|1}$ the mathematical expectation of Y,

$$\mu_{2|0} = \mathsf{E}[x - m_{1|0}]^2 = \Sigma p_{i|}[x_i - m_{1|0}]^2 = \sigma_x^2$$

the variance of X, $\mu_{0|2} = \sigma_y^2$ the variance of Y.

We shall continuously use these three systems of parameters. If the joint frequency-distribution is given, they are all uniquely determinable. On the other hand, the joint frequency-distribution is determinable uniquely in a case when a suitable choice and sufficient number of the parameters m are given or when, besides the values of $m_{1|0}$ and $m_{0|1}$, a sufficient number of the parameters μ are given, or, finally, if besides the values of $m_{1|0}$, $m_{0|1}$, $\mu_{2|0}$ and $\mu_{0|2}$ the parameters r are given. I cannot enter into the proof of these theorems here. In the case of discontinuous joint frequency-distribution it offers no particular difficulties, but in the case of continuous ones it is rather difficult to formulate. This task—known in the literature as *Problême des moments*—would take us too far afield from our principal task. Moreover, it is not of great relevance to our immediate purpose.

Among the three systems of parameters, the parameters r are distinguished by being abstract numbers ; this makes them specially well fitted for the comparison of frequency-distributions. The first in the series of r-parameters is the well-known ' coefficient of correlation '

$$r_{1|1} = \frac{\mu_{1|1}}{\mu_{2|0}^{\frac{1}{2}}\mu_{0|2}^{\frac{1}{2}}} = \frac{\mu_{1|1}}{\sigma_x\sigma_y}.$$

It can easily be shown that the absolute value of the correlation coefficient cannot be greater than 1. As the mathematical expectation of a variable which takes no negative values cannot be negative,

$$\mathsf{E}\left\{\frac{x - m_{1|0}}{\sigma_x} - r_{1|1}\frac{y - m_{0|1}}{\sigma_y}\right\}^2 \geq 0.$$

Since $$\mathrm{E}\frac{[x - m_{1|0}]^2}{\sigma_x^2} = \mathrm{E}\frac{[y - m_{0|1}]^2}{\sigma_y^2} = 1$$

and $$\mathrm{E}\frac{[x - m_{1|0}]\,[y - m_{0|1}]}{\sigma_x \sigma_y} = r_{1|1},$$

we have

$$\mathrm{E}\left\{\frac{x - m_{1|0}}{\sigma_x} - r_{1|1}\frac{y - m_{0|1}}{\sigma_y}\right\}^2 = 1 - r_{1|1}^2.$$

Hence : $$1 - r_{1|1}^2 \geq 0 \text{ and } r_{1|1}^2 \leq 1.$$

The parameters which characterize the conditional frequency-distributions will be specified by putting the relevant value of X in brackets above : thus, for instance, $m_{|1}^{(i)}$ denotes the conditional mathematical expectation of Y which corresponds to the value X_i of X, and $\mu_{2}^{(i)}$ indicates the corresponding conditional variance of Y. Thus we have according to definition :

$$m_{|1}^{(i)} = \sum_j p_{j|}^{(i)} y_j, \quad \mu_{|2}^{(i)} = \sum_j p_{j|}^{(i)} [y_j - m_{|1}^{(i)}]^2.$$

In a similar way the parameters of the conditional frequency-distributions of X are denoted by

$$m_{1|}^{(j)} = \sum_i p_{i|}^{(j)} x_i, \quad \mu_{2|}^{(j)} = \sum_i p_{i|}^{(j)} [x_i - m_{1|}^{(i)}]^2.$$

Of the many relations between the conditional and non-conditional parameters we will only note those which connect the non-conditional mathematical expectations of variables with the mean values of the conditional mathematical expectations, and further, those which connect the non-conditional variances with the mean values of conditional variances, as we shall often make use of these relations. From the definitions the following identities can be derived :

$$\sum_i p_{i|} m_{|1}^{(i)} = m_{0|1},$$

$$\sum_i p_{i|} \mu_{|2}^{(i)} = \sum_i \sum_j p_{i|} p_{j|}^{(i)} [(y_j - m_{0|1}) - (m_{|1}^{(i)} - m_{0|1})]^2 =$$

$$= \mu_{0|2} - \sum_i p_{i|} [m_{|1}^{(i)} - m_{0|1}]^2$$

$$\sum p_{|i} m_{1|}^{(j)} = m_{1|0}, \quad \sum p_{|j} \mu_{2|}^{(j)} = \mu_{2|0} - \sum_j p_{j|} [m_{1|}^{(j)} - m_{1|0}]^2.$$

Now, if the variables are mutually independent, then one has for any value of h

$$m^{(j)}_{h|} = m_{h|0} \qquad \text{for } j = 1, 2, \ldots, l$$
$$m^{(i)}_{|h} = m_{0|h} \qquad \text{for } i = 1, 2, \ldots, k$$

and also for any value of f and of g

$$m_{f|g} = m_{f|0} m_{0|g} \qquad \mu_{f|g} = \mu_{f|0}\mu_{0|g} \qquad r_{f|g} = r_{f|0} r_{0|g}.$$

Since $r_{1|0} = r_{0|1} = 0$, in this case $r_{1|1} = 0$.

It may also be proved that conversely the variables must be independent, when

$$m_{f|g} = m_{f|0} m_{0|g} \quad \text{or} \quad \mu_{f|g} = \mu_{f|0}\mu_{0|g} \quad \text{or} \quad r_{f|g} = r_{f|0} r_{0|g}$$

for all and even only for all positive integral values of f and of g.

2. As a matter of course, the conditional mathematical expectation comes first under consideration among the parameters which characterize the conditional frequency-distributions. If the relationship between the conditional mathematical expectation of Y and the corresponding values of X is expressed in analytical form, one uses $m^{(i)}_{1|} = f(x_i)$ to denote this equation, as mentioned above (Chap. III, § 4, 2): $m^{(i)}_{1|} = f(x_i)$ is the *regression equation* of Y on X and $m^{(j)}_{1|} = F(y_j)$ is the regression equation of X on Y.

The term ' regression ' originally had a definite sense in this connexion, that was lost later, so that nowadays it is considered a conventional technical term from which all etymological reminiscences should be kept at a distance.

A. In a case where the regression of Y on X takes the form of a parabola with the equation

$$m^{(i)}_{1|} = a_{|0} + a_{|1}x_i + a_{|2}x_i^2 + \ldots + a_{|f}x_i^f,$$

the coefficients of this equation can be easily expressed by the parameters m, μ, and r. It is only necessary to multiply both the sides by $p_{i|}x_i^h$, where h is at first left undetermined, and to sum for i; as $\sum_i p_{i|}x_i^h m^{(i)}_{1|} = m_{h|1}$ one arrives at

$$m_{h|1} = a_{|0}m_{h|0} + a_{|1}m_{h+1|0} + \ldots + a_{|f}m_{h+f|0}.$$

Putting in this equation h equal to 0, 1, 2, . . ., &c., up

to f, one obtains $f + 1$ linear equations, from which the $f + 1$ coefficients a are easily obtained in the form of determinants. If we then insert these coefficients in the general equation again with an undetermined value of h we obtain an equation of condition to which the parameters m must satisfy in order that the regression of Y on X may take the shape of a parabola on the fth degree.

In the case of linear regression with the equation

$$m'_{|1} = a_{|0} + a_{|1}x_i$$

we find: $a_{|1} = \dfrac{m_{1|1} - m_{1|0}m_{0|1}}{m_{2|0} - m^2_{1|0}}$ $\qquad a_{|0} = m_{0|1} - a_{|1}m_{1|0}$,

so that the equation can be written in the form

$$m'_{|1} - m_{0|1} = \frac{m_{1|1} - m_{1|0}m_{0|1}}{m_{2|0} - m^2_{1|0}}\,[x_i - m_{1|0}].$$

The condition which the parameters m have to fulfil in order that the regression of Y upon X may be linear, can be put in the form

$$\frac{m_{h|1} - m_{h|0}m_{0|1}}{m_{h+1|0} - m_{h|0}m_{1|0}} = \frac{m_{1|1} - m_{1|0}m_{0|1}}{m_{2|0} - m^2_{1|0}}$$

for any positive integral value of h.

B. The regression equation can also be expressed in another form, more convenient for many purposes. Instead of expressing the conditional mathematical expectation of Y as a function of X, the deviation of the conditional mathematical expectation from the non-conditional mathematical expectation of Y is expressed as a function of the deviation of the corresponding value of X from the mathematical expectation of X. Thus the regression equation here takes the form:

$$m'_{|1} - m_{0|1} = b_{|0} + b_{|1}[x_i - m_{1|0}] + b_{|2}[x_i - m_{1|0}]^2 + \\ + \ldots + b_{f}\,[x_i - m_{1|0}]^f.$$

If we multiply both the sides by $p_{i|}[x_i - m_{1|0}]^h$, sum for i, and notice that $\sum_i p_{i|}[x_i - m_{1|0}]^h\ [m'_{|1} - m_{0|1}] = \mu_{h|1}$, we obtain, as before, an equation which for any value of h connects the coefficients b with the parameters μ:

$$\mu_{h|1} = b_{|0}\mu_{h|0} + b_{|1}\mu_{h+1|0} + b_{|2}\mu_{h+2|0} + \ldots + b_{|f}\mu_{h+f|0}.$$

Putting h equal to 0, 1, 2, &c., up to f, we arrive at a system of $f + 1$ linear equations from which the coefficients b can be ascertained in the form of determinants. By the insertion of their values in the original general equation we obtain a new form of the condition that the regression should take the form of a parabola of the fth order.

In the case of a linear regression with the equation

$$m^{(i)}_{|1} - m_{1|0} = b_{|0} + b_{|1}[x_i - m_{1|0}]$$

we obtain $b_{|0} = 0$, $b_{|1} = \frac{\mu_{1|1}}{\mu_{2|0}}$. Hence the equation can be expressed in the form:

$$m^{(i)}_{|1} - m_{0|1} = \frac{\mu_{1|1}}{\mu_{2|0}}\,[x_i - m_{1|0}].$$

The coefficient $b_{|1} = \frac{\mu_{1|1}}{\mu_{2|0}} = \frac{\sigma_y}{\sigma_x} r_{1|1}$ in the linear equation is termed '*coefficient of regression*'—a rather curious term as, strictly speaking, all coefficients in a regression equation of any form can with equal justification be termed coefficients of regression. This meaning of the technical term 'coefficient of regression' has, however, become firmly embedded in statistical literature. If one speaks briefly of 'coefficient of regression' the coefficient $b_{|1} = \frac{\sigma_y}{\sigma_x} r_{1|1}$ in the linear regression equation is always meant.

The conditional equation which the parameters μ must satisfy in order that the regression of Y on X be linear has a simpler form than those which connect the parameters m: we must have $\frac{\mu_{h|1}}{\mu_{h+1|0}} = \frac{\mu_{1|1}}{\mu_{2|0}}$ for all positive integral values of h.

In a similar way we find when the regression of X on Y is linear,

$$m^{(j)}_{1|} - m_{1|0} = b_{1|}[y_j - m_{0|1}] = \frac{\mu_{1|1}}{\mu_{0|2}}\,[y_j - m_{0|1}]$$

and $$\frac{\mu_{1|h}}{\mu_{0|h+1}} = \frac{\mu_{1|1}}{\mu_{0|2}} \text{ at } h = 2,\ 3,\ 4,\ \ldots.$$

If both the regressions are linear one has simultaneously $b_{|1} = \frac{\sigma_y}{\sigma_x} r_{1|1}$ and $b_{1|} = \frac{\sigma_x}{\sigma_y} r_{1|1}$. Hence it follows that

$$b_{1|} b_{|1} = r_{1|1}^2 = \frac{\mu_{1|1}^2}{\mu_{2|0} \mu_{0|2}} :$$

the product of coefficients of regression $b_{1|}$ and $b_{|1}$ is identically equal to the square of the correlation coefficient, or, in other words, the correlation coefficient is equal to the geometric mean of the two coefficients of regression $b_{1|}$ and $b_{|1}$.

C. Putting in the linear regression equation of Y on X $\frac{\sigma_y}{\sigma_x} r_{1|1}$, the regression equation then takes the shape

$$m_{|1}^{(i)} - m_{0|1} = \frac{\sigma_y}{\sigma_x} r_{1|1} [x_i - m_{1|0}].$$

Now, proceeding to so-called '*normal co-ordinates*', by dividing the differences on the right and left hand by the corresponding standard deviations and by putting

$$\frac{m_{|1}^{(i)} - m_{0|1}}{\sigma_y} = \mathfrak{M}_{|1}^{(i)}, \quad \frac{x_i - m_{1|0}}{\sigma} = \mathfrak{X}_i$$

the regression equation takes a form which is particularly convenient for algebraic calculations—namely,

$$\mathfrak{M}_{|1}^{(i)} = r_{1|1} \mathfrak{X}_i.$$

Hence the conditions which the parameters r must satisfy in order that the regression of Y on X be linear—namely,

$$r_{h|1} = r_{1|1} r_{h+1|0} \text{ at } h = 2, 3, 4, \ldots$$

follow directly.

The transition to normal co-ordinates transforms all coefficients in the regression equations into abstract numbers—therein lies the advantage of this form of expression. Without working out all the calculations, I will give the regression equations in normal co-ordinates for the case

when the regression takes the form of a parabola of the second degree :

$$\mathfrak{M}_{1|1}^{(i)} = -\frac{r_{2|1} - r_{3|0}r_{1|1}}{r_{4|0} - r_{3|0}^2 - 1} + \left\{r_{1|1} - \frac{r_{3|0}[r_{2|1} - r_{3|0}r_{1|1}]}{r_{4|0} - r_{3|0}^2 - 1}\right\}\mathfrak{X}_i + \\ + \frac{r_{2|1} - r_{3|0}r_{1|1}}{r_{4|0} - r_{3|0}^2 - 1}\mathfrak{X}_i^2.$$

It is to be borne in mind that in this case the term independent of $\mathfrak{X}_i$ does not disappear in the equation, and further that when $r_{1|1} = 0$ the regression equation is reduced to

$$\mathfrak{M}_{1|1}^{(i)} = \frac{r_{2|1}}{r_{4|0} - r_{3|0}^2 - 1}\left\{\mathfrak{X}_i^2 - r_{3|0}\,\mathfrak{X}_i - 1\right\}.$$

We shall have an opportunity later on to return to these differences from linear regression.

The equation of condition that the regression of Y on X can be expressed by a parabola of the second degree can be put in the form :

$$\frac{r_{h|1} - r_{1|1}\,r_{h+1|0}}{r_{h+2|0} - r_{h+1|0}r_{3|0} - r_{h|0}} = \frac{r_{2|1} - r_{1|1}r_{3|0}}{r_{4|0} - r_{3|0}^2 - 1} \text{ for } h = 3,\ 4,\ 5,\ \ldots$$

3. When the variable Y is not correlated with the variable X, in the sense of Karl Pearson's definition (cf. Chap. III, § 4, 3), it is if the conditional mathematical expectation of Y remains the same for all values of X, the regression equation of Y on X reduces to $m_{1|1}^{(i)} - m_{0|1} = 0$, that is $\mathfrak{M}_{1|1}^{(i)} = 0$, respectively.

In order that the variable Y be uncorrelated with X, the parameters m, μ, and r must satisfy the conditions :

$$m_{h|1} = m_{h|0}m_{0|1},\ \mu_{h|1} = 0,\ r_{h|1} = 0 \text{ for } h = 1,\ 2,\ 3,\ \ldots$$

Putting in $r_{h|1} = 0$ $h = 1$, we arrive at $r_{1|1} = 0$. Thus the non-correlation of the variable Y with X implies that the correlation coefficient equals zero. It does not follow, however, that conversely in a case where the coefficient of correlation is zero, that variables are uncorrelated : from $r_{1|1} = 0$ one may infer the non-correlation only if it is certain that the regression is linear. For with linear regres-

sion $m_{|1}^{(i)} = m_{0|1}$ follows directly from $r_{1|1} = 0$. If, on the contrary, the regression is non-linear, it does not by any means follow from $r_{1|1} = 0$ that $m_{|1}^{(i)} = m_{0|1}$. We have observed that in the example of parabolic regression. This must be carefully borne in mind.

One must remember, further, that from $\mu_{h|1} = 0$ for $h = 1, 2, \ldots$ it can be inferred that $\mu_{1|h} = 0$ at $h = 2, 3, \ldots$: from the non-correlation of the variable Y with X it cannot be concluded that the variable X is uncorrelated with Y.

If variables are mutually independent they are also uncorrelated. This follows directly from the definitions of the notions and can also be gathered from the equations of condition; as $\mu_{1|0} = \mu_{0|1} = 0$ we obtain from $\mu_{f|g} = \mu_{f|0}\mu_{0|g}$ both $\mu_{f|1} = 0$ and $\mu_{1|g} = 0$ for all values of f and g.

However, the non-correlation does not imply mutual independence—not even when Y is uncorrelated with X as well as X with Y: it can be concluded from $\mu_{h|1} = 0$ and $\mu_{1|h} = 0$ for $h = 1, 2, 3, \ldots$ that $\mu_{f|g} = \mu_{f|0}\mu_{0|g}$ when f and g are different from 1.

4. Suppose that the true regression of Y on X takes the shape of a parabola of the fth degree, and let us try to fit a straight line, so that it is the best representation of the true regression, in the sense that the sum of the squares of the deviations of the conditional mathematical expectations calculated from the equation of the straight line from the corresponding true values of the conditional mathematical expectations of Y is less than that given by any other line. If the equation of the required straight line is written in the form $M_{|1}^{(i)} = A_{|0} + A_{|1}x_i$, then we have to choose coefficients $A_{|0}$ and $A_{|1}$, so that

$$\sum_i p_{i|}[m_{|1}^{(i)} - M_{|1}^{(i)}]^2 = \sum_i p_{i|}[m_{|1}^{(i)} - A_{|0} - A_{|1}x_i]^2$$

has its minimum value. Bearing in mind that

$$\sum_i p_{i|} = 1, \quad \sum_i p_{i|}x_i = m_{1|0}, \quad \sum_i p_{i|}m_{|1}^{(i)} = m_{0|1}, \quad \sum_i p_{i|}x_i m_{|1}^{(i)} = m_{1|1},$$

$$\sum_i p_{i|}x_i^2 = m_{2|0},$$

we arrive at the equations

$m_{0|1} - A_{|0} - A_{|1}m_{1|0} = 0$ and $m_{1|1} - A_{|0}m_{1|0} - A_{|1}m_{2|0} = 0$,

from which the values of $A_{|0}$ and $A_{|1}$ can be determined as

$$A_{|1} = \frac{m_{1|1} - m_{1|0}m_{0|1}}{m_{2|0} - m_{1|0}^2} = \frac{\mu_{1|1}}{\mu_{2|0}},\quad A_{|0} = m_{0|1} - A_{|1}m_{1|0} = \frac{m_{2|0}m_{0|1} - m_{1|1}m_{1|0}}{m_{2|0} - m_{1|0}^2}.$$

Thus we obtain for $A_{|0}$ and $A_{|1}$ the same values as if the true regression were a linear one (cf. above, § 3, 2, A). The straight line for which the equation is

$$m'_{|1} - m_{0|1} = \frac{\mu_{1|1}}{\mu_{2|0}}[x_i - m_{1|0}] \text{ or } \mathfrak{M}'_{|1} = r_{1|1}\mathfrak{X}_i$$

is thus the true regression line in the case where the true regression of Y on X is linear and represents the true regression line with the best approximation in the case where the true regression departs from linearity. This property of the straight lines $\mathfrak{M}'_{|1} = r_{1|1}\mathfrak{X}_i$ is of great value to the statistician. He is under all circumstances able to gain a fairly appropriate idea of the regression by calculating the equation of the straight line $\mathfrak{M}'_{|1} = r_{1|1}\mathfrak{X}_i$.

5. The regression equation elucidates one of those properties of the joint frequency-distribution which awaken the inquirer's greatest interest. The importance of the regression equation to the inquirer who has to deal with stochastically associated variables is analogous to that of the law in a case of functional relationship. The knowledge of the formula of the law enables us to infer from the given value of X value of Y without using direct measurements of Y. In a similar way the regression equation conveys knowledge of the expected value of the variable Y, which corresponds to each given value of the variable X. The possibility of calculating with certainty beforehand the value which the variable Y will obtain, it is true, is not given ; for in the case of stochastic association the variable Y retains the character of a chance variable even after the determination of the value of X and its possible values

fluctuate after the determination of the value of X round the conditional mathematical expectation, as they do round the non-conditional mathematical expectation before the determination of the value of X. But the measure of the fluctuation which is given by the standard deviation or by the variance is considerably reduced. Setting aside the regression equation and considering the variable Y apart from its relationship with X one has to reckon with the variance $\mu_{0|2}$. However, if one is able to refer to the regression equation and, for each given value of X, proceed from the corresponding conditional mathematical expectation $m_{|1}^{(i)}$, then as a measure of the fluctuation which corresponds to the value X_i of X we have the conditional variance $\mu_{|2}^{(i)}$. However, because, as we have seen,

$$\sum_i p_{i|}\mu_{|2}^{(i)} = \mu_{0|2} - \sum_i p_{i|}[m_{|1}^{(i)} - m_{0|1}]^2,$$

so the mean size of the fluctuations of the possible residual values of Y round the conditional mathematical expectation of Y after the value of X has been determined, i.e. $\sum_i p_{i|}\mu_{|2}^{(i)}$, is smaller than $\mu_{0|2}$, except in the case where $m_{|1}^{(i)} = m_{0|1}$ at $i = 1, 2, \ldots, k$, and the variable Y is uncorrelated with X. Thus, in a case of non-correlation we gain nothing if in considering the variable Y we proceed from the value which X has taken. In all other cases, however, if the regression equation is known, the knowledge of the value obtained for X is of importance, as it enables us to calculate beforehand with less uncertainty the value of Y which must be expected.

6. As with the conditional mathematical expectation, one can deal in a similar way with other coefficients which characterize conditional frequency-distributions: conditional variance, or conditional standard deviations, various measures of asymmetry of the frequency-distribution, &c. They are presented as functions of the X-values; the coefficients of the equations in question can be expressed by means of the parameters m, μ, and r; conditions can be deduced that the equations in question should take a defi-

nite form, linear, &c. Theoretically, those problems are of great interest for the development of methods which aim at examining stochastic association between two or more variables. In order not to overburden the presentation I shall not go into the matter any further. Besides, we can dispense with detailed treatment the more easily as the methods in question do not greatly contribute to novelty of notion, and are meanwhile applied relatively seldom in practice.

§ 4

1. A third main group of methods aimed at by the comprehensive examination of the joint frequency-distributions is the calculation of coefficients which resemble mean square contingency inasmuch as they are designed to express numerically certain features of the stochastic association of the variables whose values are under examination; but which, again, differ from mean square contingency in so far as they apply, not only the probabilities of the possible values of the variables but also use these possible values themselves. We must examine two such coefficients more closely, the correlation coefficient and the so-called correlation ratio. The calculation of the correlation coefficient appears nowadays to be the most popular of the methods which can be applied in investigating two stochastically associated variables. However, as the importance of the correlation coefficient to the inquirer is partly founded on the fact that the correlation coefficient, under certain circumstances, is equal to the correlation ratio, let us begin with the consideration of the latter.

2. I have already pointed out (cf. § 3, 5) how useful is the mean conditional variance $\Sigma_i p_{i|} \mu_{|2}^{(i)}$ as characteristic of stochastic association between variables. The coefficient which has been devised by Karl Pearson* and called by

* K. Pearson, on the general theory of skew correlation and non-linear regression, p. 10 (Drapers' Company Research Memoirs, Biometric Series, II; 1905).

him '*correlation ratio*' and denoted by the Greek letter η is just based on the magnitude of the mean conditional variance. The correlation ratio of Y on X is defined by the relation

$$\eta^2_{y|x} = 1 - \frac{1}{\mu_{0|2}} \sum_i p_{i|} \mu^{(i)}_{|2}.$$

Consequently the correlation ratio is nothing else than the amount to be added to the quotient of the mean conditional variance by the total non-conditional variance to make unity. As we know (cf. above, § 3, 1),

$$\sum_i p_{i|} \mu^{(i)}_{|2} = \mu_{0|2} - \sum_i p_{i|} [m^{(i)}_{|1} - m_{0|1}]^2,$$

it follows from the definition that

$$\eta^2_{y|x} = \frac{1}{\mu_{0|2}} \sum_i p_{i|} [m^{(i)}_{|1} - m_{0|1}]^2,$$

or in normal co-ordinates is (cf. above, § 3, 2, C)

$$\eta^2_{y|x} = \sum_i p_{i|} [\mathfrak{M}^{(i)}_{|1}]^2.$$

The property of stochastic association of the variable Y with the variable X consists in the fact that Y remains a chance variable which can assume different values with definite probabilities even when the value of the variable X is fixed; in other words, the conditional variances of the variable Y remain different from zero. If Y stands in functional relationship to X, then all conditional variances are equal to zero and the correlation ratio of Y on X then equals 1. Conversely, if the correlation ratio of Y on X is equal to 1, this implies that the mean conditional variance of Y is equal to zero, which can be the case only if all separate conditional variances vanish, that is; if the relationship is functional. As the mean conditional variance cannot be negative, the value of the correlation ratio can never be greater than 1. This greatest possible value of the correlation ratio thus characterizes the presence of a functional relationship between the variable Y and the variable X and implies that the expected value of Y, after that of X has been fixed, can be predicted with certainty. Again,

the value of the correlation ratio can never decrease below zero, as the mean conditional variance can never be greater than the non-conditional total variance; indeed, $\sum_i p_{i|}\mu_{|2}^{(i)}$ is equal, as we have seen, to $\mu_{0|2} - \sum_i p_{i|}[m_{|1}^{(i)} - m_{0|1}]^2$. The correlation ratio can equal zero only when the mean conditional variance equals the total variance, hence when $\sum_i p_{i|}[m_{|1}^{(i)} - m_{0|1}]^2 = 0$. This necessitates that all magnitudes $m_{|1}^{(i)}$ are equal among themselves, that thus Y and X are uncorrelated. Conversely, the correlation ratio is equal to zero when all magnitudes $m_{|1}^{(i)}$ are equal and Y is uncorrelated to X. In this case the knowledge of the value of X gives, as we have already pointed out, no advantage in the respect of the prediction of the expected value of Y. The values of the correlation ratio which lie between these extremes can be intelligibly interpreted likewise. The greater the correlation ratio the more considerably reduced will appear the mean conditional variance in comparison with the total variance, and the more closely the prediction of the value of Y to be expected, for any value of X will approach an estimate made on the basis of a formula expressing in functional form the relationship between Y and X. The smaller the value of the correlation ratio the greater is the range of chance fluctuations to which the determination of the value of Y is exposed after the determination of the value of X—the more uncertain will be our estimate of the expected value of Y on a basis of the knowledge of the regression equation and the value of X.

The correlation ratio of Y on X gives the intensity of the association of the variable Y with the variable X a numerical expression which keeps the same meaning for any joint frequency-distribution. Herein lies an essential advantage over the correlation coefficient, the numerical values of which (cf. below, § 4, 3) have the meaning which is usually attributed to them only in the case of linear regression. As an absolute measure of intensity the correlation ratio does not hold good either. From the value 0

of the correlation ratio of Y on X can be deduced the non-correlation but not the range of fluctuation of which Y remains capable after the value of X has been fixed. If the correlation ratio of Y on X is exactly 0, it means that the determination of the value of the variable X, on the average, leaves the range of fluctuations of Y unchanged; but we gain no information from the value of the correlation ratio itself as to whether this range is great or small: rather it must be supplemented by the value of $\sum_i p_{i|}\mu^{(i)}_{|2} = \mu_{0|2}[1 - \eta^2_{y|x}]$ if the absolute intensity of the relationship is to be estimated in this sense. If the association of Y with X is homoscedastic, then the constant conditional variance of Y which then appears in the place of the mean conditional variance of Y, is identically equal to $\mu_{0|2}[1 - \eta^2_{y|x}]$.

In the calculation of the correlation ratio we can start from $\eta^2_{y|x} = 1 - \frac{1}{\mu_{0|2}}\sum_i p_{i|}\mu^{(i)}_{|2}$, as well as from $\eta^2_{y|x} = \sum_i p_{i|}[\mathfrak{M}^{(i)}_{|1}]^2$. If the regression equation is known, then the correlation ratio can be expressed, by means of substituting the value of $m^{(i)}_{|1}$, in terms of the coefficients of the regression equation and by means in terms of the parameters m, μ, and r. Thus we obtain, for instance, if the regression takes the form of a parabola of the second degree, from the equation known to us (cf. above, § 3, 2, C)

$$\mathfrak{M}^{(i)}_{|1} = -\frac{r_{2|1} - r_{3|0}r_{1|1}}{r_{4|0} - r^2_{3|0} - 1} + \left\{r_{1|1} - \frac{r_{3|0}[r_{2|1} - r_{3|0}r_{1|1}]}{r_{4|0} - r^2_{3|0} - 1}\right\}\mathfrak{X}_i + \\ + \frac{r_{2|1} - r_{3|0}r_{1|1}}{r_{4|0} - r^2_{3|0} - 1}\mathfrak{X}^2_i,$$

after some transformations

$$\eta^2_{y|x} = r^2_{1|1} + \frac{[r_{2|1} - r_{3|0}r_{1|1}]^2}{r_{4|0} - r^2_{3|0} - 1}.$$

3. If the regression of Y on X is linear with the equation $\mathfrak{M}^{(i)}_{|1} = r_{1|1}\mathfrak{X}_i$, we obtain, since $\sum_i p_{i|}\mathfrak{X}^2_i = 1$, $\eta^2_{y|x} = r^2_{1|1}$. Hence in a case of a linear regression the absolute magni-

tudes of the correlation ratio and the correlation coefficient are equal.

If, however, the regression is curvilinear, then $r^2_{1|1} < \eta^2_{y|x}$. The easiest way of proving this is as follows. Let us denote by $M^{(i)}_{|1} = m_{0|1} + \frac{\mu_{1|1}}{\mu_{2|0}}[x_i - m_{1|0}]$ the equation of the straight line best fitting the true line of regression. Since $\sum_i p_{i|}[m^{(i)}_{|1} - M^{(i)}_{|1}]^2 = \mu_{0|2}[\eta^2_{y|x} - r^2_{1|1}]$ we have

$$\eta^2_{y|x} - r^2_{1|1} = \frac{1}{\mu_{0|2}} \sum_i p_{i|}[m^{(i)}_{|1} - M^{(i)}_{|1}]^2.$$

Obviously the difference $\eta^2_{y|x} - r^2_{1|1}$ cannot be negative. It can equal zero only, when all magnitudes $m^{(i)}_{|1}$ coincide with the corresponding magnitudes $M^{(i)}_{|1}$, i.e. when the straight line $M^{(i)}_{|1} = m_{0|1} + \frac{\mu_{1|1}}{\mu_{2|0}}[x_i - m_{1|0}]$ coincides with the true line of regression.

Hence, if the regression is linear, numerical values of the correlation coefficient are to be interpreted in exactly the same way as the values of the correlation ratio coincident with them. If the correlation coefficient is equal to 1, then the variable Y stands in linear functional relationship to X. If the correlation coefficient is equal to 0, then the variable Y is uncorrelated with X. The greater the value of the correlation coefficient the narrower is the range of chance fluctuations to which the variables Y remain exposed after the value of X has been fixed—the more closely the association between Y and X approximates to a linear functional relationship. However, if the regression is not linear, one must no longer interpret the numerical values of the correlation coefficient in the above sense. In a curvilinear regression the correlation coefficient is always smaller in absolute magnitude than the correlation ratio. A value of the correlation coefficient smaller than 1 corresponds to the value 1 of the correlation ratio: hence, a functional non-linear relationship between Y and X corresponds not to a value 1 of the correlation coefficient but to a smaller one, more or less divergent from the value 1 according

to the particular kind of functional relationship. Again, the value of 0 of the correlation coefficient in a case where the regression is non-linear does not imply that the variable Y is uncorrelated to X. Thus both the limits 0 and 1, for non-linear regression, have not the same sense in which they are to be interpreted when the regression is linear. If one does not know in advance whether the regression is linear, one must not infer from the value 0 of the correlation coefficient that the variables are uncorrelated. It is just as inadmissible to conclude, when the correlation coefficient remains under 1, that there is no functional relationship; it is not excluded that the relationship is nevertheless functional but non-linear.

Hence, if one is not sure whether the regression is linear, one must be very careful in interpreting the numerical values of the correlation coefficient. In order to avoid misinterpretation in such cases it is advisable to calculate the values of the correlation ratio which retain the same meaning in any law of dependence. Therefore in measuring the intensity of association the correlation ratio must be preferred to the correlation coefficient. The correlation coefficient may be used as a correct measure of intensity only when it can represent the correlation ratio, as then the numerical values of the two coincide. Otherwise, the intensity of an association would be systematically underestimated if it were measured by means of the correlation coefficient.

Yet the interest of the value of the correlation coefficient is not confined to the fact that the correlation coefficient can, with certain reservations, be an appropriate measure of the intensity of association between variables. We have seen (cf. above, § 3, 4) that even in non-linear regression the straight line whose equation in terms of normal coordinates is $\mathfrak{M}_{|1}^{(i)} = r_{1|1}\mathfrak{X}_i$ may hold good as an approximate expression of the true line of regression, and as such always offers a certain interest to an inquirer—even if it is only a kind of a preliminary reconnaissance of the field concerned.

This line shows us whether the conditional expected value of Y increases or decreases on the average with the increase of X, and gives as well the average size of the increase or decrease. The equation of this line is determined by means of the value of the correlation coefficient. If the correlation coefficient is positive, the conditional mathematical expectation of Y increases the more markedly with the growth of X, the greater the correlation coefficient ; if the correlation coefficient is negative, the decrease of the mathematical expectation becomes more marked with increasing X, the more closely the correlation coefficient tends to -1. Again, the equation of the straight line which reproduces the best fit for the true regression of X on Y is likewise determined by the value of the correlation coefficient $r_{1|1}$. Thus the value of the correlation coefficient gives us, in summary form, a fair amount of information on the association between the variables even in a case where the regression is non-linear.

Sometimes, the numerical value of the correlation coefficient in linear regression permits of particularly meaningful interpretations. Let us perhaps assume that one throws m white and n red dice and one denotes by U the sum thrown by the red dice, by W that thrown by the white dice and one puts $X = U + W$. Letting the white dice lie, one picks up the red ones, shakes them in the dice-box and throws again ; one denotes by T the sum shown by throwing the red dice the second time and one puts $Y = W + T$. The correlation coefficient between X and Y is easily calculated under such circumstances and is equal to the ratio between the number of the common addenda —i.e. m—and the total number of addenda—$m + n$: i.e. $r_{1|1} = \frac{m}{m+n}$. Thus one can infer from the value of the correlation coefficient the relative number of white dice which still lay on the table. The scheme of this example can be generalized : when both the variables are represented as sums of mutually independent addenda, which follow

the same law of distribution, then the correlation coefficient is equal to the ratio of the number of the common addenda to the geometric mean of the number of addenda in the two sums: $r_{1|1}=\frac{m}{\sqrt{[m+n]\,[m+l]}}$, if m denotes the number of common addenda, $m+n$ the total number of addenda in X, and $m+l$ the total number in Y.

§ 5

1. Particularly great importance is attached to the correlation coefficient when the stochastic association between variables takes the form of so-called '*normal correlation*'. One understands by 'normal correlation' the case where the variables X and Y can take continuously all values between $-\infty$ and $+\infty$, and the probability of the coincidents of a value of X which lies between x and $x+dx$ with a value of Y which lies between y and $y+dy$ is equal to

$$\frac{1}{2\pi\sqrt{\mu_{2|0}\mu_{0|2}-\mu_{1|1}^2}}\, e^{-\frac{\mu_{0|2}[x-m_{1|0}]^2-2\mu_{1|1}[x-m_{1|0}]\,[y-m_{0|1}]+\mu_{2|0}[y-m_{0|1}]^2}{2[\mu_{2|0}\mu_{0|2}-\mu_{1|1}^2]}}dxdy.$$

If we put $\mathfrak{X}=\frac{x-m_{1|0}}{\sqrt{\mu_{2|0}}}$ $\mathfrak{Y}=\frac{y-m_{0|1}}{\sqrt{\mu_{0|2}}}$, we obtain as the probability of the coincidents of values of $\mathfrak{x}$ which lie between $\mathfrak{X}$ and $\mathfrak{X}+d\mathfrak{X}$ with those of $\mathfrak{Y}$ which lie between $\mathfrak{Y}$ and $\mathfrak{Y}+d\mathfrak{Y}$ the expression, in terms of normal co-ordinates:

$$\frac{1}{2\pi\sqrt{1-r_{1|1}^2}}e^{-\frac{\mathfrak{X}^2-2r_{1|1}\mathfrak{X}\mathfrak{Y}+\mathfrak{Y}^2}{2[1-r_{1|1}^2]}}d\mathfrak{X}\,d\mathfrak{Y}.$$

Hence the law of dependence is determined, in the case of normal correlation, by the value of the correlation coefficient conjointly with the values of both the mathematical expectations and both the variances $m_{1|0}$, $\mu_{2|0}$, $m_{0|1}$, $\mu_{0|2}$, which characterize the frequency-distributions of X and Y,

and serve to determine the system of normal co-ordinates. Consequently the value of the correlation coefficient appears in the case of normal correlation to be the key to all requisite knowledge required of the stochastic association of variables : if one knows the value of $r_{1|1}$, then the law of dependence is ascertained and all the rest is deducible by normal mathematical operations.

I shall not enter into the closer consideration of these formal mathematical deductions. They are tasks of integration which offer no difficulties in the case of two variables. I shall only convey some results which describe more exactly the content of the notion of ' normal correlation ' which is so important to the theory of statistics.

2. As the value of the correlation coefficient $r_{1|1}$ determines the law of dependence, all other indices, in the case of normal correlation, can be presented as functions of $r_{1|1}$.

We obtain for the higher r-parameters, the following values :

$$r_{3|1} = r_{1|3} = 3r_{1|1} \qquad r_{2|2} = 1 + 2r_{1|1}^2 \qquad r_{4|0} = r_{0|4} = 3$$

$$r_{5|1} = r_{1|5} = 15r_{1|1} \qquad r_{4|2} = r_{2|4} = 3[1 + 4r_{1|1}^2]$$

$$r_{3|3} = 3r_{1|1}[3 + 2r_{1|1}^2] \qquad r_{6|0} = r_{0|6} = 15 \qquad r_{f|2h+1-f} = 0$$

$$r_{2f|2h} = 1\cdot3\cdot5 \ldots (2f-1)\cdot1\cdot3\cdot5 \ldots (2h-1)\left\{1 + \sum_{t=1} \frac{2^{2t}\, f^{t-}\, h^{t-}}{1\cdot2\cdot3 \ldots (2t)} r_{1|1}^{2t}\right\}$$

$$r_{2f+1|2h+1} = 1\cdot3\cdot5 \ldots (2f+1)\cdot1\cdot3\cdot5 \ldots (2h+1) r_{1|1} \sum_{t=0} \frac{2^{2t}\, f^{t-}\, h^{t-}}{1\cdot2\cdot3 \ldots (2t+1)} r_{1|1}^{2t}$$

where $\qquad z^{t-} = z(z-1)\,(z-2) \ldots (z-t+1).$

The law of dependence is symmetrical. The frequency-distributions of X and of Y take the form of the Gauss-Laplace law of error. The regression of Y on X, as well as that of X on Y is linear. Consequently the correlation coefficient and both the correlation ratios are identically equal, and the correlation coefficient may be considered as

a reliable measure of the intensity of association. If the correlation coefficient is equal to zero, then

$$r_{2f+1|2h+1} = 0 = r_{2f+1|0}r_{0|2h+1},$$

$$r_{2f|2h} = 1\cdot3\cdot5 \ldots (2f-1)\cdot1\cdot3\cdot5 \ldots (2h-1) = r_{2f|0}r_{0|2h}$$

and the variables are accordingly (cf. above, § 3, 1) mutually independent.

All conditional frequency-distributions of X as well as of Y likewise take the form of the Gauss-Laplace's law of errors. The dependence of the variable X on Y, as well as that of the variable Y on X is homoscedastic. The conditional variance of X remains constant for all values of Y and has the value $\mu_{2|}^{(j)} = \mu_{2|0}[1 - r_{1|1}^2]$. The conditional variance of Y retains for all values of X the same value $\mu_{|2}^{(i)} = \mu_{0|2}[1 - r_{1|1}^2]$.

It is interesting to point out further that the mean square contingency is associated in the case of normal correlation with the correlation coefficient by the equation $\varphi^2 = \frac{r_{1|1}^2}{1 - r_{1|1}^2}$, whence $r_{1|1} = \pm\sqrt{\frac{\varphi^2}{1+\varphi^2}}$.

3. The preference given to normal correlation in the theory of statistics is due partly to historical causes. The modern theory of correlation was at first even more closely attached to normal correlation, than is the theory of methods whose aim is the examination of a chance variable, to the Gauss-Laplace's law of error. The notions ' correlation coefficient ', ' regression equation ', &c., are based on the consideration of normal correlation, and its ' shells ' remained attached to them for a very long time. It was not until the end of the nineteenth century, for example, that people began to distinguish in the correlation coefficient those properties which it has in the case of normal correlation from those which it has when the regression is linear, as well as from those when the law of dependence is of any form whatever. Nowadays, thanks in the first place to Karl Pearson and G. Udny Yule, we realize that normal

correlation only is one of the possible forms of the law of dependence. But about a quarter of a century ago we were unable to think at all clearly about the stochastic association of chance variables in any other form than this. This former monopoly helps normal correlation even nowadays to obtain a prominent position not only in our textbooks but also in our theoretical systems.

But there are more important grounds for giving this prominent position to normal correlation. We must first of all take into consideration the fact that the mathematical analysis is considerably simplified and more elegant in shape, all higher r-parameters being eliminated from the general formulae by the assumption of normal correlation as they can all be expressed in normal correlation as functions of the coefficient of correlation: by this means the formulae not only gain clarity of arrangement but are more manageable in practical application. This relative simplicity of mathematical and computational treatment, together with the fact that the development of the modern theory of correlation has concentrated almost entirely on the consideration of normal correlation for so long, has resulted in theory of normal correlation being now relatively complete: many tasks which we shall soon have under consideration are already, for the case of normal correlation, more or less satisfactorily solved, while their investigation for the general case of any law of dependence is not yet so far advanced; frequently it has hardly even been tackled. Therefore the theory of normal correlation can be presented with a completeness and finish as yet unachieved by the general case.

As regards the actual occurrence of normally correlated chance variables, we are not indeed inclined any longer to assume that normal correlation is very often to be found or even as a general rule; but cases with no excessive deviation from normal correlation are not infrequent, so in view of its relative simplicity we often adopt methods of inquiry which rest on the assumption of normal correlation. It is

of importance here that, as in the case of the Gauss-Laplace's law of error, the association between the means of empirical values of chance variables asymptotically approaches normal correlation with increasing size of sample for almost any form of the law of dependence between the variables provided the separate samples are independent. For two greatly non-normal laws of dependence between the variables the approach is even rather rapid, so that in practice, in almost all cases where two stochastically associated means are involved, one makes use of the assumption of the normal correlation which simplifies the treatment.

§ 6

The form of the regression equation of Y on X and the intensity of the association between Y and X are not mutually dependent. Whether Y is functionally related to X or whether it is quite loosely associated with X is entirely irrelevant to the form of the regression line. When Y is functionally related to X, then all conditional variances of Y are equal to zero, the correlation coefficient of Y on X is equal to 1, and the regression line of Y on X graphically represents the law of functional relationship ; hence it can, according to circumstances, take the form of a straight line or of a curve for any degree of complication. Hence, one cannot infer from the shape of the regression line of Y on X whether there is a greater or smaller intensity of an association.

Just as little inference with regard to the intensity of an association can be made from the form of the regression line of X on Y. But the simultaneous consideration of both the regression lines permits us to gain a certain insight into the intensity of the association.

When X and Y are functionally related, the equation which expresses Y as an explicit function of X is derivable from the equation expressing X as an explicit function of Y. If one proceeds from the consideration of one regression line, one is always able to ascertain how the other regres-

sion line would appear if the relationship between the variables were a functional one. Thus, if the actual regression of the second variable upon the first one is of a different form from the other, one may consider the assumption that the variables then in functional relationship to each other is refuted : the variables are then stochastically associated.

Let us assume that the regression of Y on X is linear and that the regression equation takes the form $m_{|1}^{(i)} = a_{|0} + a_{|1}x_i$. When both the variables are functionally related $m_{1|}^{(i)}$ coincides with those values of Y which correspond to the value X_i of X as again $m_{1|}^{(j)}$ coincides with those values of X which correspond to the value Y_j of Y. Obviously the regression equation of Y on X on the assumption of functional relationship may be written in the form $y = a_{|0} + a_{|1}x$. Hence, we obtain : $x = -\frac{a_{|0}}{a_{|1}} + \frac{1}{a_{|1}}y$. In a case where there is a functional relationship between the variables, the regression equation of X on Y must accordingly be presented in the shape $m_{1|}^{(i)} = -\frac{a_{|0}}{a_{|1}} + \frac{1}{a_{|1}}y_j$. Thus if the line of regression of X on Y is not straight, or if, in the case of linear regression of X on Y, in the regression equation the coefficient $a_{1|}$ is different from $\frac{1}{a_{|1}}$, then the presence of a functional relationship between X and Y is out of the question.

It can be decided in a similar way whether or not the assumption of functional relationship can be justified in cases where neither regression is linear.

When both the regressions are linear we are in a position to estimate exactly the intensity of association between X and Y from the two regression equations. If the regression equations take the form

$$m_{|1}^{(i)} = a_{|0} + a_{|1}x_i \text{ and } m_{1|}^{(j)} = a_{0|} + a_{1|}y_j$$

the product of the coefficients $a_{|1}$ and $a_{1|}$ is identically equal to the square of the correlation coefficient (cf. above, § 3, 2) : $a_{1|}a_{|1} = r_{1|1}^2$. If the coefficients $a_{1|}$ and $a_{|1}$ are known, the value of the correlation coefficient can be easily calculated.

But since in a case where both the regressions are linear the square of the correlation coefficient is identically equal to the correlation ratios $\eta^2_{y|x}$ and $\eta^2_{x|y}$ (cf. above, § 4, 3), the product of the coefficients $a_{1|}$ and $a_{|1}$ in this case gives that measure of intensity of association between the variables which we have recognized as being the best one.

When the actual regression lines of Y on X and of X on Y are non-linear, but the equations of the straight lines giving the best fit (cf. above, § 3, 4) are known, we are able likewise to determine the value of the correlation coefficient since the product of the coefficients $A_{1|}$ and $A_{|1}$ in the equations of these lines is likewise identically equal to the square of the correlation coefficient. However, under such circumstances we are no longer entitled to infer the values of the correlation ratios from the value of the correlation coefficient. When the actual line of regression of Y on X is non-linear, the correlation coefficient is in absolute magnitude always smaller than the correlation ratio of Y on X (cf. above, § 4, 3). Under such circumstances the numerical value of the correlation coefficient which we calculate as the geometric mean of the values of the coefficients $A_{1|}$ and $A_{|1}$ no longer appears as an exact measure of the actual intensity of the association between X and Y. When one makes an estimate from the numerical value of the correlation coefficient, the intensity is then more or less underrated, according to the actual form of the regression lines. The correlation coefficient remains less than 1 even when the relationship between X and Y is functional.

§ 7

Our survey of the methods of examining two stochastically associated variables is far from being exhausted. I have had to confine myself to the systematic development of the fundamental conceptions upon which the modern theory of the methods applied by statisticians rests. We cannot enter into the description of its detailed formation, of its adaptation to the special peculiarities of particular prob-

lems, nor can we enter into details in the form of the statistician's material. We cannot even consider in greater detail the attractive problem of the statistical investigation of chance variable—non-quantitative characteristics which we touched upon in our consideration of mean square contingency. However, I must not fail to mention that the notions of the correlation coefficient and of the correlation ratio, although in a general case they proceed from the supposition of quantitatively different values of a chance-variable magnitude, may be applied in the investigation of the association of non-quantitative chance-variable characteristics if the characteristics each permit of only two different categories. In order to see this in the first place one must assume that the variable magnitude X as well as Y can assume only two different numerical and constant values, and on this assumption one must calculate the correlation coefficient between X and Y according to the general formula $r_{1|1} = \frac{\mu_{1|1}}{\sqrt{\mu_{2|0}\mu_{0|2}}}$. If, as before, we denote (cf. above, § 2, 1) by δ the difference $p_{1|1}p_{2|2} - p_{1|2}p_{2|1}$ we easily find

$$\mu_{1|1} = \delta[x_1 - x_2]\,[y_1 - y_2],\ \mu_{2|0} = p_{1|}p_{2|}[x_1 - x_2]^2,$$
$$\mu_{0|2} = p_{|1}p_{|2}[y_1 - y_2]^2.$$

The numerical values of the variables thus appear in the numerator and denominator of the correlation coefficient in the form of products $[x_1 - x_2]\,[y_1 - y_2]$ and can be reduced so that for the correlation coefficient between X and Y we obtain the value

$$r_{1|1} = \frac{\delta}{\sqrt{p_{1|}p_{2|}p_{|1}p_{|2}}}.$$

When the variables can each assume only two different values, the formula obviously does not contain the possible values of the variables. The correlation coefficient remains unchanged when the possible values of the variables change arbitrarily, provided the probabilities of the values remain the same. Accordingly, it is unnecessary to know the pos-

sible values in order to calculate the correlation coefficient ; hence it is unnecessary to measure them ; it is even unnecessary to assume that they are measurable at all.

As we see, the value of the correlation coefficient which we have obtained is equal to that at which we arrive for the mean square contingency (cf. above, § 2, 1). The calculation of both the correlation ratios leads likewise to the same expression. The magnitude $\frac{\delta^2}{p_{|1}p_{2|}p_{|1}p_{|2}}$ can be regarded in a case where the chance-variable characteristics permit of only two different categories as the mean square contingency φ^2, as well as the square of the correlation coefficient $r_{1|1}$, or as either of the correlation ratios $\eta^2_{y|x}$ and $\eta^2_{x|y}$.

CHAPTER V

THE EMPIRICAL MATERIAL AND THE COEFFICIENTS WHICH SUMMARIZE IT

§ 1

THE law of dependence and the complete set of parameters and coefficients which summarize it provide a knowledge of stochastic association between chance variables sufficient for all purposes. The investigation of stochastic association always aims in the first place at ascertaining as reliably as possible the numerical values of those of these magnitudes which the inquirer decides on, in the case concerned, for objective or practical reasons. Only after this problem has been so more or less satisfactorily solved in a ' mathematical ' or ' elementary ' form can the final test be undertaken, the elucidation of the true meaning of the relationships determined (cf. Chap. II, § 2, and Chap. VIII, § 2).

In practice, the investigator is rarely able to set up the law of dependence by means of deductions from the theory of probability so as to elucidate its general properties by examples taken from the realm of so-called games of chance. As a rule, what is known to the inquirer about the variables to be investigated and their mutual relations does not go further than the knowledge of several pairs of corresponding chance values of both the variables ; the *a priori* magnitudes interesting to the inquirer have to be estimated on the basis of those chance values of variables. To lead the way from the empirical chance values of the variables to the required *a priori* magnitudes is part of the main object of the theory of correlation. However, before turning to the systematic survey of the methods concerned we must take

closer cognizance of the empirical material which is to be used here and to consider how it can be thrown into the forms best suited to further use.

§ 2

1. The material at the inquirer's disposal consists of a number of pairs of corresponding chance variables of X and Y. Let us write: N—the total number of pairs; $n_{i|}$—the number of pairs in which X has the value X_i; $n_{|j}$—the number of pairs in which Y has the value Y_j; $n_{i|j}$—the number of pairs in which X has the value X_i and Y the value Y_j. If we denote by k the number of different values of X and by l those of Y, then

$$n_{i|} = \sum_{j=1}^{l} n_{i|} \qquad n_{|j} = \sum_{i=1}^{k} n_{i|j}$$

$$N = \sum_{i=1}^{k}\sum_{j=1}^{l} n_{i|j} = \sum_{i=1}^{k} n_{i|} = \sum_{j=1}^{l} n_{|j}.$$

If the numbers $n_{i|j}$ are put into the clearly arranged form of correlation table considered in Table 1, Chapter I, one usually terms the horizontal rows as well as the vertical columns of the table 'arrays', while the separate arrays are characterized by the value of the variable which remains constant for all members of the array. The total of the numbers $n_{i|j}$ for constant i forms the X_i array; the total of the numbers $n_{i|j}$ for constant j forms the Y_j array. If we put

$$\frac{n_{i|j}}{N} = p'_{i|j} \qquad \frac{n_{i|}}{N} = p'_{i|} \qquad \frac{n_{|j}}{N} = p'_{|j}$$

$$\frac{n_{i|j}}{n_{i|}} = \frac{p'_{i|j}}{p'_{i|}} = p^{(i)'}_{|j} \qquad \frac{n_{i|j}}{n_{|j}} = \frac{p'_{i|j}}{p'_{|j}} = p^{(j)'}_{i|},$$

then the set of numbers $p'_{i|}$ gives the empirical frequency-distribution of X and the set of numbers $p'_{|j}$ that of Y. Similarly, $p^{(i)}_{|j}$ give the empirical frequency-distribution of the values which fall respectively in the X_i and Y_j arrays of the variable placed in the Y_j array.

Empirical Material and Coefficients

Let us now consider the pairs of corresponding values of X and Y as to be numbered and let us denote by $X^{[f]'}$ and by $Y^{[f]'}$ the values of the fth pair. Let us further denote by X'_0 the arithmetic mean of all the X values, by $X_0^{(j)}$ that of the X values in the array Y_j, by Y'_0 the arithmetic mean of all values of Y and by $Y_0^{(i)'}$ that of the Y-values in the array X_i. From the definitions we obtain the identities:

$$x'_0 = \frac{1}{N}\sum_{f=1}^{N} x^{[f]'} = \frac{1}{N}\sum_{i=1}^{k} n_{i|}x_i = \sum_{i=1}^{k} p'_{i|}x_i$$

$$y'_0 = \frac{1}{N}\sum_{f=1}^{N} y^{[f]'} = \frac{1}{N}\sum_{j=1}^{l} n_{|j}y_j = \sum_{j=1}^{l} p'_{|j}y$$

$$x_0^{(j)'} = \frac{1}{n_{|j}}\sum_{i=1}^{k} n_{i|j}x_i = \sum_{i=1}^{k} p_{i|}^{(j)'}x_i$$

$$y_0^{(i)'} = \frac{1}{n_{i|}}\sum_{j=1}^{l} n_{i|j}y_j = \sum_{j=1}^{l} p_{|j}^{(i)'}y_j.$$

Since

$$\sum_{j=1}^{l} n_{|j}x_0^{(j)'} = \sum_{j=1}^{l}\sum_{i=1}^{k} n_{i|j}x_i = \sum_{i=1}^{k} n_{i|}x_i,$$

we obtain further

$$x'_0 = \frac{1}{N}\sum_{i=1}^{k} n_{i|}x_i = \frac{1}{N}\sum_{j=1} n_{|j}x_0^{(j)'} = \sum_{j=1} p'_{|j}x_0^{(j)'}.$$

Similarly we arrive at

$$y'_0 = \sum_{i=1}^{k} p'_{i|}y_0^{(i)'}.$$

2. The *a priori* law of dependence may, as we have seen, be taken as given, provided that the parameters $m_{f|g}$ are known (cf. Chap. IV, § 3, 1). In a similar way the set of numbers $p'_{i|j}$ can be expressed by parameters $m'_{f|g}$ defined by the relations:

$$m'_{f|g} = \sum_i\sum_j p'_{i|j}x_i^f y_j^g.$$

Putting $g = 0$ we obtain parameters $m'_{f|0}$ which express

the frequency-distribution of X. Putting $f = 0$ we obtain parameters $m'_{f|g}$, which express the frequency-distribution of Y. Thus the parameter

$$m'_{1|0} = \sum_i \sum_j p'_{i|j} x_i = \sum_i p'_{i|} x_i = x'_0$$

denotes the arithmetic mean of all the X-values and the parameter

$$m'_{0|1} = \sum_j p'_{|j} y_j = y'_0$$

this of all Y-values.

The set of numbers $p'_{i|j}$ can also be expressed by the parameters $\mu'_{f|g}$ and $r'_{f|g}$, which are defined by the relations

$$\mu'_{f|g} = \sum_i \sum_j p'_{i|j} [x_i - m'_{1|0}]^f [y_j - m'_{0|1}]^g$$

$$r'_{f|g} = \frac{\mu'_{f|g}}{[\mu'_{2|0}]^{\frac{1}{2}f} [\mu'_{0|2}]^{\frac{1}{2}g}}.$$

Putting here $g = 0$ (and $f = 0$ respectively) we obtain likewise the parameters which express the frequency-distributions of the values of X (and those of Y respectively). Thus

$$\mu'_{2|0} = \sum_i p'_{i|} [x_i - m'_{1|0}]^2 = \frac{1}{N} \sum_i n_{i|} [x_i - x'_0]^2 = m'_{2|0} - [m'_{1|0}]^2$$

gives the *empirical variance* of X and

$$\mu'_{0|2} = \sum_j p'_{|j} [y_j - m'_{0|1}]^2 = \frac{1}{N} \sum_j n_{|j} [y_j - y'_0]^2 = m'_{0|2} - [m'_{0|1}]^2$$

that of Y.

Let us call the first of the series of r'-parameters, viz.

$$r'_{1|1} = \frac{\mu'_{1|1}}{\sqrt{\mu'_{2|0}\mu'_{0|2}}} = \frac{\sum_i \sum_j p'_{i|j} [x_i - m'_{1|0}] [y_j - m'_{0|1}]}{\sqrt{\{\sum_i p'_{i|} [x_i - m'_{1|0}]^2\} \{\sum_j p'_{|j} [y_j - m'_{0|1}]^2\}}} =$$

$$= \frac{\sum_{f=1}^{N} [x^{[f]\prime} - x'_0][y^{[f]\prime} - y'_0]}{\sqrt{\left\{\sum_{f=1}^{N} [x^{[f]\prime} - x'_0]^2\right\} \left\{\sum_{f=1}^{N} [y^{[f]} - y'_0]^2\right\}}},$$

the *empirical correlation coefficient*. Since

$$\mu'_{1|1} = m'_{1|1} - m'_{1|0} m'_{0|1},$$

the empirical correlation coefficient can also be represented by the form

$$r'_{1|1} = \frac{m'_{1|1} - m'_{1|0}m'_{0|1}}{\sqrt{\{m'_{2|0} - [m'_{1|0}]^2\}\{m'_{0|2} - [m'_{0|1}]^2\}}}.$$

It can easily be shown that the empirical correlation coefficient cannot be greater in absolute value than 1. Since the sum of magnitudes which are not negative cannot be negative, we have

$$\sum_i \sum_j p'_{i|j}\left[\frac{x_i - m'_{1|0}}{\sqrt{\mu'_{2|0}}} - r'_{1|1}\frac{y_i - m_{0|1}}{\sqrt{\mu'_{0|2}}}\right]^2 \geq 0.$$

Since again

$$\sum_i \sum_j p'_{i|j}\left[\frac{x_i - m'_{1|0}}{\sqrt{\mu'_{2|0}}}\right]^2 = \frac{\sum_i p'_{i|}[x_i - m'_{1|0}]^2}{\mu'_{2|0}} = 1,$$

$$\sum_i \sum_j p'_{i|j}\left[\frac{y_i - m'_{0|1}}{\sqrt{\mu'_{0|2}}}\right]^2 = 1, \quad \sum_i \sum_j p'_{i|j}\frac{[x_i - m'_{1|0}]}{\sqrt{\mu'_{2|0}}}\frac{[y_j - m'_{0|1}]}{\sqrt{\mu'_{0|2}}} = r'_{1|1},$$

we have $1 - 2[r'_{1|1}]^2 + [r'_{1|1}]^2 \geq 0$ or $[r'_{1|1}]^2 \leq 1$.

3. As before (cf. Chap. IV, § 3, 1), let us distinguish the parameters referring to the separate arrays from those which refer to the marginal totals of the respective variable by adding in brackets on top, on the right, a reference to the particular array. Putting, accordingly,

$$m^{(i)'}_{|g} = \sum_j p^{(i)'}_{|j} y^g_j \qquad m^{(j)'}_{f|} = \sum_i p^{(j)'}_{i|} x^f_i$$

$$\mu^{(i)'}_{|g} = \sum_j p^{(i)'}_{|j}[y_j - m^{(i)'}_{1|}]^g \qquad \mu^{(j)'}_{f|} = \sum_j p^{(j)'}_{i|}[x_i - m^{(j)'}_{1|}]^f,$$

$m^{(j)'}_{1|}$ and $m^{(i)'}_{1|}$ denote the arithmetic means of the values of X for the Y_j-array and of Y for the X_i-array, and $\mu^{(j)'}_{2|}$ and $\mu^{(i)'}_{2|}$ respectively, the corresponding variances.

Of the identities which connect these parameters, we shall bear in mind only those to which we shall have frequently to refer, viz.:

$$m'_{1|0} = \sum_j p'_{|j} m^{(j)'}_{1|} \qquad m'_{0|1} = \sum_i p'_{i|} m^{(i)'}_{1|}$$

$$\mu'_{2|0} = \sum_j p'_{|j}\mu^{(j)'}_{2|} + \sum_j p'_{|j}[m^{(j)'}_{1|} - m'_{1|0}]^2$$

$$\mu'_{0|2} = \sum_i p'_{i|}\mu^{(i)'}_{2|} + \sum_i p'_{i|}[m^{(i)'}_{1|} - m'_{0|1}]^2.$$

§ 3

If the values of $m_{11}^{(i)\prime}$, which correspond to different values of X, be graphically represented by a rectangular co-ordinate system and the successive points are connected by straight lines, we then obtain a broken line which we denote as *empirical regression line* of Y on X.

In considering the empirical regression line a problem may be raised similar to that we examined when considering the *a priori* regression line, viz. (cf. Chap. I, § 3, 2): under the assumption that the separate points $m_{11}^{(i)\prime}$ of the irregular empirical regression line lie on a curve which has the form of a parabola of the fth degree, to express the coefficients of the equation of the parabola by the parameters m', μ', and r'. The mathematical treatment of this problem is not formally different from that of § 3 of Chapter IV; but the task does not offer any greater statistical interest in the case of the empirical regression line. Even when the corresponding values of the variables X and Y are actually so constituted that all the points $m_{11}^{(i)\prime}$ lie exactly on a parabola of not too high a degree, the statistician is unable to make any inferences of value from this: we must always bear in mind that this may be due to chance, and adding of further pairs of empirical values of X and Y may remove the seeming simplicity of the course of the line.

When considering the empirical regression line there is much greater interest in finding the equations of the straight line or curve of relatively simple form which may best represent the set of points considered in the sense of the Method of Least Square, while we must reckon from the outset with the fact that not all of the separate points lie on the curve, but some of them are grouped irregularly round it, at greater or smaller distances. The problem is therefore to plot a straight line in such a way that a weighted mean of squares of differences between the actual mean values of all the individual X_i arrays and the correspond-

ing values calculated according to the equation of the line is as small as possible (cf. Chap. IV, § 3, 4). If we write the equation of the line in the form

$$M'^{(i)}_{|1} = A'_{|0} + A'_{|1}x_i,$$

the condition that the sum

$$\sum_i p'_{i|}[m'^{(i)}_{|1} - M'^{(i)}_{|1}]^2 = \sum_i p'_{i|}[m'^{(i)}_{|1} - A'_{|0} - A'_{|1}x_i]^2$$

is as small as possible leads to equations

$$m'_{0|1} - A'_{|0} - m'_{1|0}A'_{|1} = 0, \quad m'_{1|1} - m'_{1|0}A'_{|0} - m'_{2|0}A'_{|1} = 0.$$

Hence we obtain:

$$A'_{|1} = \frac{m'_{1|1} - m'_{1|0}m'_{0|1}}{m'_{2|0} - [m'_{1|0}]^2} = \sqrt{\frac{\mu'_{0|2}}{\mu'_{2|0}}}r'_{1|1}, \quad A'_{|0} = m'_{0|1} - A'_{|1}m'_{1|0} =$$

$$= \frac{m'_{2|0}m'_{0|1} - m'_{1|1}m'_{1|0}}{m'_{2|0} - [m'_{1|0}]^2}.$$

Accordingly, the equation of the curve has the form

$$M'^{(i)}_{|1} - m'_{0|1} = r'_{1|1}\sqrt{\frac{\mu'_{0|2}}{\mu'_{2|0}}}[x_i - m'_{1|0}].$$

We shall call the coefficient of $[x_i - m'_{1|0}]$ in the equation of the line which best represents the empirical regression line of Y on X in the sense defined above the '*empirical regression coefficient* of Y on X' and denote it by $b'_{|1}$ (cf. Chap. IV, § 3, 2): consequently we have according to the definition:

$$b'_{|1} = r'_{1|1}\sqrt{\frac{\mu'_{0|2}}{\mu'_{2|0}}}.$$

Similarly we find for the curve which best represents the empirical regression line of X on Y the equation:

$$M'^{(j)}_{1|} - m'_{1|0} = r'_{1|1}\sqrt{\frac{\mu'_{2|0}}{\mu'_{0|2}}}[y_j - m'_{0|1}];$$

accordingly, the empirical regression coefficient of X on Y is

$$b'_{1|} = r'_{1|1}\sqrt{\frac{\mu'_{2|0}}{\mu'_{0|2}}}.$$

Thus the empirical correlation coefficient $r'_{1|1}$ is equal to the geometric mean of both the empirical regression coefficients (cf. Chap. IV, § 3, 2):

$$r'_{1|1} = \sqrt{b'_{1|}b'_{|1}}.$$

§ 4

Let $$[\eta'_{y|x}]^2 = 1 - \frac{1}{\mu'_{0|2}} \sum_i p'_{i|} \mu^{(i)'}_{|2}.$$

We shall call the magnitude $\eta'_{y|x}$ '*the empirical correlation ratio* of Y on X' (cf. Chap. IV, § 4). None of the magnitudes $\mu^{(i)'}_{|2}$ can be negative; consequently neither can their weighted mean be negative. Hence $[\eta'_{y|x}]^2$ cannot exceed 1.

By the substitution (cf. above, § 2, 3)

$$\sum_i p'_{i|} \mu^{(i)'}_{|2} = \mu'_{0|2} - \sum_i p'_{i|} [m^{(i)'}_{|1} - m'_{0|1}]^2$$

we obtain

$$[\eta'_{y|x}]^2 = \frac{1}{\mu'_{0|2}} \sum_i p'_{i|} [m^{(i)'}_{|1} - m'_{0|1}]^2.$$

Since none of the magnitudes $p'_{i|} [m^{(i)'}_{|1} - m'_{0|1}]^2$ can be negative also, $[\eta'_{y|x}]^2$ cannot be negative. Thus the numerical value of the empirical correlation ratio lies between 0 and 1:

$$0 \leq [\eta'_{y|x}]^2_i \leq 1.$$

The empirical correlation ratio of Y on X is equal to 1 when all magnitudes $\mu^{(i)'}_{|2}$ are equal to zero. In order that all magnitudes $\mu^{(i)'}_{|2}$ shall be equal to zero, it is necessary and sufficient that all the values of Y which correspond to each one of those X-values are equal. Let us assume that the number of pairs of corresponding X and Y is equal to the number of different values of X. Then to every value of X the only one value of Y corresponds; in this case all magnitudes $\mu^{(i)}_{|2}$ are equal to zero; the empirical correlation ratio of Y on X becomes identically equal to 1. If the empirical material is so constituted no safe inference can be made from the value 1 of the empirical correlation ratio with regard to the connexion between Y and X.

The empirical correlation ratio is equal to zero, when

$$\mu'_{0|2} = \sum_i p'_{i|} \mu^{(i)'}_{|2} \text{ or } \sum_i p'_{i|} [m^{(i)'}_{|1} - m'_{0|1}]^2 = 0,$$

i.e. when all magnitudes $m^{(i)'}_{|1}$ are by chance exactly equal. It is well to remember here that the empirical magnitudes

$m_{|1}^{(i)\prime}$ can be equal to each other without *a priori* magnitudes $m_{|1}^{(i)}$ being equal to one another.

It does not follow from $[\eta'_{y|x}]^2 = 0$ that the variable Y is uncorrelated with the variable X (cf. Chap. IV, § 4): the magnitudes $m_{|1}^{(i)\prime}$ can be equal to each other because of their chance deviations from the corresponding values $m_{|1}^{(i)}$, although the latter show a more or less considerable variance.

While retaining the notation of § 3, we easily obtain

$$[\eta'_{y|x}]^2 - [r'_{1|1}]^2 = \frac{1}{\mu'_{0|2}} \sum_i p'_i [m_{|1}^{(i)\prime} - M_{|1}^{(i)\prime}]^2.$$

Obviously the empirical correlation coefficient cannot exceed in absolute value the empirical correlation ratio (cf. Chap. IV, § 4). The empirical correlation coefficient can equal the empirical correlation ratio of Y on X only, if all points $m_{|1}^{(i)\prime}$ lie on the straight line with the equation

$$M_{|1}^{(i)\prime} - m'_{0|1} = r'_{1|1} \sqrt{\frac{\mu'_{0|2}}{\mu'_{2|0}}} [x_i - m'_{1|0}].$$

Here one must also bear in mind that the separate points $m_{|1}^{(i)\prime}$ of the empirical regression line can lie on a straight line without the regression of Y on X being necessarily linear. There is nothing to prevent the true regression equation of Y on X, which connects the values of the *a priori* magnitudes $m_{|1}^{(i)}$ with the values of X, from taking a different form, while the chance deviations of the values $m_{|1}^{(i)\prime}$ from the corresponding values $m_{|1}^{(i)}$ create an appearance of a linear regression. If the difference $\eta^2_{y|x} - r^2_{1|1}$ is equal to zero it obviously follows that the regression of Y on X is linear. If, on the contrary, the difference $[\eta'_{y|x}]^2 - [r'_{1|1}]^2$ equals zero, it only follows that the assumption that the true regression of Y on X is linear only holds good more or less plausibly.

§ 5

The empirical correlation coefficient $r'_{1|1}$ which is defined by the relation

$$r'_{1|1} = \frac{\mu'_{1|1}}{\sqrt{\mu'_{2|0}\mu'_{0|2}}}$$

can exceed, in absolute value, neither the empirical correlation ratio of Y on X nor that of X on Y (cf. above, § 4):

$$[r'_{1|1}]^2 \leq [\eta'_{y|x}]^2, \qquad [r'_{1|1}]^2 \leq [\eta'_{x|y}]^2.$$

Since the empirical correlation ratio cannot itself be greater than 1 (cf. above, § 4) we have (cf. above, § 2, 2)

$$[r'_{1|1}]^2 \leq 1.$$

That the numerical value of the empirical correlation coefficient must lie between -1 and $+1$ can also be shown in the following way. From

$$\sum_i \sum_j p'_{i|j} \left[\frac{x_i - m'_{1|0}}{\sqrt{\mu'_{2|0}}} - \frac{y_j - m'_{0|1}}{\sqrt{\mu'_{0|2}}} \right]^2 \geq 0$$

and

$$\sum_i \sum_j p'_{i|j} \frac{[x_i - m'_{1|0}]^2}{\mu'_{2|0}} = \sum_i \sum_j p'_{i|j} \frac{[y_j - m'_{0|1}]^2}{\mu'_{0|2}} = 1$$

$$\sum_i \sum_j p'_{i|j} \frac{[x_i - m'_{1|0}]\,[y_j - m'_{0|1}]}{\sqrt{\mu'_{2|0}\mu'_{0|2}}} = r'_{1|1}$$

we find $1 - 2r'_{1|1} + 1 \geq 0$. Hence $r'_{1|1} \leq +1$.

Again it follows from

$$\sum_i \sum_j p'_{i|j} \left[\frac{x_i - m'_{1|0}}{\sqrt{\mu'_{2|0}}} + \frac{y_j - m'_{0|1}}{\sqrt{\mu'_{0|2}}} \right]^2 \geq 0$$

in a similar way that $1 + 2r'_{1|1} + 1 \geq 0$, $r'_{1|1} \geq -1$.

The empirical correlation coefficient can have the value $+1$ only if all differences $\frac{x_i - m'_{1|0}}{\sqrt{\mu'_{2|0}}} - \frac{y_j - m'_{0|1}}{\sqrt{\mu'_{0|2}}}$ are equal to 0, i.e. if only one value of y corresponds to each value of x and the deviations of the corresponding values of the variables from their respective mean values are exactly proportional to each other and always have the same sign. The empirical correlation coefficient can have the value -1 only if all sums $\frac{x_i - m'_{1|0}}{\sqrt{\mu'_{2|0}}} + \frac{y_j - m'_{0|1}}{\sqrt{\mu'_{0|2}}}$ are equal to 0, i.e. if only one value of y corresponds to each value of x and the deviations of the corresponding values of variables from their mean are exactly proportional to each other and

always have opposite signs. From $r'_{1|1} = \pm 1$ it follows accordingly that the empirical regression line is linear. However, it does not imply without any further consideration that the true regression is linear. It is not impossible that the appearance of a linearity is counterfeited by chance deviations of $m'^{(i)}_{1|1}$ values from the corresponding $m^{(i)}_{1|1}$ values.

The empirical correlation coefficient can become indeterminate. If all empirical values of X or all empirical values of Y are by chance equal, then the magnitude $\mu'_{1|1}$ which occurs in the numerator of the empirical correlation coefficient, as well as one of the variances in the denominator, is equal to zero. Thus the empirical correlation coefficient assumes the value $\frac{0}{0}$. When the set of pairs of corresponding values of the variables at our disposal is small, one must always reckon with the possibility that the empirical correlation coefficient becomes indeterminate.

§ 6

Let

$$[\varphi']^2 = \sum_i \sum_j \frac{[p'_{i|j} - p'_{i|}p'_{|j}]^2}{p'_{i|}p'_{|j}} = \sum_i \sum_j \frac{\left[n_{i|j} - \frac{1}{N} n_{i|} n_{|j}\right]^2}{n_{i|} n_{|j}}.$$

We shall denote the magnitude φ' as the *empirical mean square contingency* (cf. Chap. IV, § 2, 1). From

$$\sum_i \sum_j n_{i|j} = N, \quad \sum_i \sum_j n_{i|} n_{|j} = \left[\sum_i n_{i|}\right] \left[\sum_j n_{|j}\right] = N^2$$

we have

$$[\varphi']^2 = \sum_i \sum_j \frac{n^2_{i|j}}{n_{i|} n_{|j}} - 1.$$

The empirical mean square contingency is equal to zero if all differences $p'_{i|j} - p'_{i|}p'_{|j}$ are equal to zero. However, it does not follow from here without any further consideration that all differences $p_{i|j} - p_{i|}p_{|j}$ are also equal to zero and the variables X and Y are mutually independent (cf. Chap. IV, § 2, 1): the mutual independence might be counterfeited by chance. If all empirical values of the

variable Y which correspond to each value of the variable X are equal, then $p'_{i|j} = 0$ for $j \neq i$ and again $p'_{i|j}$ is equal to $p'_{i|}$. For $[\varphi']^2$ we then obtain the value $k - 1$. Here the assumption that the variables are functionally related seems to be more or less plausible. But in this case, also, we must bear in mind the reservations which were pointed out above, when considering the value 1 of the empirical correlation ratio (cf. above, § 4).

§ 7

Under the assumption that both the variables X and Y have taken only two different values each, we put (cf. Chap. IV, § 2, 1)

$$p'_{1|1}p'_{2|2} - p'_{1|2}p'_{2|1} = \delta'.$$

From the identities, which connect the magnitudes $p'_{i|j}$, $p'_{i|}$, $p'_{|j}$, we easily obtain

$$\delta' = p'_{1|1} - p'_{1|}p'_{|1} = p'_{2|2} - p'_{2|}p'_{|2} = -[p'_{1|2} - p'_{1|}p'_{|2}] = \\ = -[p'_{2|1} - p'_{2|}p'_{|1}].$$

If we put these values of the differences $p'_{1|1} - p'_{1|}p'_{|1}$, &c., in the formula which defines the empirical mean square contingency (cf. above, § 6), we obtain

$$[\varphi']^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{[\delta']^2}{p'_{i|}p'_{|j}} = \frac{[\delta']^2}{p'_{1|}p'_{2|}p'_{|1}p'_{|2}}.$$

In a case where both the variables have taken only two different values each, both the empirical correlation ratios and the square of the empirical correlation coefficient can be reduced to the same expression

$$\frac{[\delta']^2}{p'_{1|}p'_{2|}p'_{|1}p'_{|2}}.$$

CHAPTER VI

ESTIMATE OF *A PRIORI* COEFFICIENTS ON THE BASIS OF EMPIRICAL MATERIAL

§ 1

WE have (cf. Chap. IV) surveyed, in outline, the methods in common use for extracting the stochastical connexion between two chance variables and so seen what it is that the statistician tries to learn by his analysis of data. We have also considered the material from which, as a rule, he has to proceed (cf. Chap. V). We may now ask how the features of stochastical connexion between the variables considered by the inquirer may be determined. The knowledge of *a priori* frequency-distributions which would allow him to calculate the numerical values of *a priori* magnitudes by means of formulae which define them without any reference to experience are at the statistician's disposal only in exceptional cases. As a rule, he only knows pairs of associated empirical values assumed by the correlated variables in a series of ' experiments '. Hence it is the task of the theory of statistics to show how it is possible to advance from these empirical chance values of variables to those numerical values which comprehensively characterize the unknown *a priori* joint frequency-distribution in a way appropriate to the object of an inquiry.

To state this problem with scientific precision and to solve it in a special case was J. Bernoulli's claim to immortality in our subject. The solution rests on the law of great numbers which links empirical numbers which are statistically determinable on the one hand with *a priori* magnitudes which are their basis but as a rule are inaccessible for

immediate determination on the other hand. The fundamental idea of the law of great numbers can be expressed in manifold forms. For our purposes the most useful way appears to proceed from the formulation which in its simplest form reduces to the well-known Cebysheff* inequality. Assume that N different experiments are carried out on a chance variable ; the probability that the deviation of the arithmetic mean of the chance values, which the variable takes at these N experiments, from the *a priori* mathematical expectation, becomes smaller than a preassigned small quantity asymtotically approaches the limit 1 with increasing N ; consequently in a fairly large number of experiments it can be assumed that the mathematical expectation of the variable and the average of its empirically chance values do not differ much from each other. We cannot enter here into the mathematical proof of Bernoulli's and of Cebysheff's theorems, nor can we dwell upon complicated logical problems involved in the law of great numbers. We must confine ourselves to the use of the law of great numbers for the construction of the theory of correlation and show with its aid how we may estimate the unknown *a priori* magnitudes determining the correlation from the empirical values of the variable given to the statistician. The law of great numbers is one of the most important fundamental pillars of the theory of correlation as well as of the general theory of statistics ; but the logical analysis of the law of great numbers is a problem of the general theory of statistics and not of the special theory of correlation.

§ 2

1. We can proceed to estimate *a priori* magnitudes from the empirical values of variable in various ways. The simplest is the following.

* About the spelling of this name see L. Isserlis, ' Note on Tchebycheff's Interpolation Formula ', *Biometrika*, Vol. 19, 1927, p. 87.—TRANS.

Estimate of A Priori *Coefficients*

In order to estimate approximately the numerical values of an *a priori* magnitude U, which can be derived by means of well-known formulae from the joint frequency-distribution, we construct a function of empirical values of variable —U'—which fulfils the condition that its mathematical expectation equals U. The numerical value of U' which follows, if the empirical values of X and Y are inserted in the formula defining U', is considered as an approximate value or, as we should rather say, as a *presumptive* value of U. In each separate case the numerical value of U' can deviate more or less from the value U; the range of such chance deviations is characterized by the magnitude of the standard deviation of U'. On the average all these chance deviations to one side or the other of U tends to zero. As a rule we shall achieve the aim of our estimations most safely if we continually keep to this method of estimation. Assume, for instance, that from a closed urn N balls are extracted, and that the white balls appear n times. The presumptive value of the unknown probability p of drawing a white ball from the urn can equal $\frac{n}{N}$ in the case where every ball extracted is replaced before the next extraction takes place as well as when it is not replaced in the urn: under both assumptions the mathematical expectation of $\frac{n}{N}$ becomes exactly equal to p for any number of extractions. Let us now repeat the experiment of N extractions from an urn t times, either always replacing the extracted ball before the next extraction takes place or not replacing the extracted balls in the urn until the conclusion of each series of N extractions. Write: n_1 for the number of white balls drawn in the first N extractions, n_2 for the number in second series, &c. According to the law of great numbers which serves us as a guiding principle there is not much difference between $\frac{1}{t}\left[\frac{n_1}{N} + \frac{n_2}{N} + \ldots + \frac{n_t}{N}\right]$ and the sought *a priori* probability p, if t be fairly large.

Hence, if we put the presumptive value p equal to the ratio of the number of white balls to the total number extracted, then though we run the risk in individual cases of our estimations being affected by a certain chance error so that the presumptive value p assumed by us amounts to something either larger or smaller than p, we shall succeed best in the long run if we make our estimates in this manner.

The calculation of the standard error of U' allows us to set limits to the range of chance errors with which individual estimates are affected. If the extracted balls are replaced in the urn before the next extraction takes place the standard error of the quota of white balls comes to what is known as

$$\sigma_{\frac{n}{N}} = \sqrt{E\left[\frac{n}{N} - p\right]^2} = \sqrt{\frac{1}{N}p(1-p)}.$$

Where the extracted balls have not been replaced in the urn it comes to

$$\sigma_{\frac{n}{N}} = \sqrt{\frac{A-N}{A-1}\frac{1}{N}p(1-p)},$$

where A denotes the totality of balls in the urn at the beginning of the experiment. If the extracted balls are not replaced the estimate is thus safer than where they are replaced in the urn before the next extraction takes place, and the reliability of the estimate grows more rapidly with the increase of N than in the first case. If $N = A$ then, in this case, $\frac{n}{N} = 0$, and the estimate is entirely reliable : when all balls are extracted from the urn we can naturally exactly infer the magnitude of the probability p from the quota of white balls.

2. Apart from this method of estimation one often tries a rougher one which, as regards the function of the empirical values of U' to be applied, is content with the demand that the mathematical expectation of U' should tend asymptotically with the increase of the number of trials,

towards the sought-for *a priori* magnitude U, and renounces the more rigorous demand that the mathematical expectation of U' should equal the *a priori* magnitude U in any finite number of trials. The numerical value of the function U' obtained after insertion of the empirical values of X and Y is considered as presumptive value of U. This manner of estimation is justified by the fact that with a fairly large number of trials the value of the mathematical expectation of U' does not deviate very much from the exactly estimated presumptive value of U, as both coincide when the number of trials becomes infinitely large. Consequently the mathematical expectation of U' may hold good as an approximate value of the true presumptive value.

This method is preferably applied in the form that, for the estimation of a function of *a priori* probabilities we make use of the same function of corresponding statistical frequencies, proceeding from the conception that the numerical values of the two functions cannot diverge too far in a fairly extensive number of trials, since, with an increasing number of trials, the frequencies tend to their respective probabilities as limits. So, for instance, we take as a presumptive value of *a priori* mean square contingency (cf. Chap. IV, § 2)

$$\varphi^2 = \sum_i \sum_j \frac{[p_{i|j} - p_{i|}p_{|j}]^2}{p_{i|}p_{|j}},$$

the numerical value of the empirical mean square contingency (cf. Chap. V, § 6)

$$[\varphi']^2 = \sum_i \sum_j \frac{[p'_{i|j} - p'_{i|}p'_{|j}]^2}{p'_{i|}p'_{|j}}.$$

Similarly the numerical value of the empirical correlation coefficient $r'_{1|1}$ (cf. Chap. V, § 2, 2) is used for the presumptive value of the *a priori* correlation coefficient $r_{1|1}$ (cf. Chap. IV, § 3, 1,) the numerical value of the empirical correlation ratio (cf. Chap. V, § 4) for the presumptive value of the *a priori* correlation ratio (cf. Chap. IV, § 4, 2), &c.

This method evidently has the disadvantage that at

repeated estimations the average of the presumptive values obtained in this way does not necessarily coincide with the true value of the corresponding *a priori* magnitude. This method of estimation is affected not only with the otherwise avoidable chance error of estimation, but also with a systematical error of estimation, which may be positive as well as negative, and thus the true value is at times systematically overestimated or systematically underestimated. When, for instance, we put the presumptive value of p^2 equal to $\left[\frac{n}{N}\right]^2$ we overestimate the sought magnitude, as the mathematical expectation of $\left[\frac{n}{N}\right]^2$ is not p^2, but $p^2 + \frac{1}{N}p(1-p)$. On the other hand, when we put the presumptive value of $p(1-p)$ equal to $\frac{n}{N}\left[1 - \frac{n}{N}\right]$, we underestimate the sought magnitude, since

$$E\frac{n}{N}\left[1 - \frac{n}{N}\right] = \frac{N-1}{N}p(1-p).$$

Since, however, $\lim_{N\to\infty}\left[p^2 + \frac{1}{N}p(1-p)\right] = p^2$

and $\lim_{N\to\infty}\left\{\frac{N-1}{N}p(1-p)\right\} = p(1-p)$,

we find in both cases that the value of the estimate reached this way really tends asymptotically, with increasing number of trials, to the exactly estimated presumptive value and that systematic errors are quite insignificant if the number of trials is fairly large.

That this also holds good as a rule has been proved in a general manner by Professor G. Bohlmann.* This method of estimation, which has been applied by statisticians mostly in an uncritical manner, has thus obtained a firmer basis.

* G. Bohlmann, Formulierung und Begründung zweier Hilfssätze der mathematischen Statistik' (*Mathematische Annalen*, Vol. 74, 1913).

If it is supplemented by an examination of the magnitude, or at least, of the sign of the systematic error, then it may be said to be scientifically correct though less satisfactory than the method first described. In the early days of the modern theory of statistics this method of inquiry rendered services which cannot be too highly praised. Most of the *a priori* magnitudes to be considered by a statistician have at first been estimated in this manner on the basis of empirical material; it was only later realized that it may bring in systematic errors; in each case we have sought to ascertain whether such are really present, and in the affirmative case to form an exact idea of their magnitude. The result of such examinations of systematic errors varied in different cases. Sometimes one was able to prove that there was no systematic error of estimation at all. In other cases one succeeded in the exact determination of the systematic error and was then able to modify the selected function U' that the systematic error was eliminated; if one puts as a basis of the estimate of the value of $p(1-p)$ instead of the magnitude $\frac{n}{N}\left(1-\frac{n}{N}\right)$ the value of $\frac{n}{N-1}\left(1-\frac{n}{N}\right)$, then the systematic error is eliminated, since the mathematical expectation of $\frac{n}{N}\left(1-\frac{n}{N}\right)\frac{N}{N-1}$ exactly equals $p(1-p)$. In very many cases, however, one could not arrive at more than an approximate estimate of systematic errors. When the empirical function U' is not an integral rational function of the values X and Y, one is as a rule forced to be content with a rough estimate of the systematic error.

3. Before we proceed to the detailed presentation of the ways in which the *a priori* magnitudes connected with correlation used to be estimated from the empirical material, let us consider more closely the notion of presumptive value. I avoid calling the presumptive values so obtained for the *a priori* magnitudes, approximate values, for they are not

approximate values in the accepted meaning : the presumptive value is a notion in itself, and approximate value *sui generis*. This comes to light clearly in that the conventional approximate value is improved by calculation to more decimal places : 3·14 is, for instance, a better approximate value of π than 3·1, and 3 would be a still worse approximate value of π. The presumptive value, on the contrary, is not improved by any increase in the number of decimal places. It can in this sense be exact without ceasing to remain approximate value in its proper sense. When, in a series of 100 extractions from a closed urn, the white ball appears 50 times, then the presumptive value $\frac{1}{2} = 0{\cdot}5$ for the *a priori* probability of the extraction of a white ball is arithmetically quite exact ; this does not, however, mean that the probability exactly equals $\frac{1}{2}$. The presumptive value is related to the *a priori* value of which an estimate is required, in the same way as a number drawn from an urn is related to the average of the numbers in the urn. The *a priori* magnitude concerned is represented by the presumptive value without being measured by it in the same sense as a constant to be estimated is measured by the usual approximate value. The measure of the exactness of the presumptive value is determined not by the number of significant figures but by the magnitude of its standard deviation. The notional peculiarity of the presumptive value consists in this, that in itself it is a chance variable which can assume different values with definite probabilities.

§ 3

1. After these introductory deliberations, let us now consider how the statistician has to proceed in the estimation of *a priori* magnitudes which comprehensively characterize the joint frequency-distribution of two stochastically associated chance variables, if he is depending exclusively on the empirical material of the frequencies with which the various combinations of possible values of X and Y have appeared

in a series of trials. We will assume here that individual trials are mutually independent, so that the probabilities $p_{i|j}$ of the coincidence of the various values of X and of Y in a trial are not influenced by the result of other trials and that the law of dependence remains the same in all N trials.

We shall also renounce here the exhaustive treatment of the subject. Our endeavour will rather be to demonstrate procedure as clearly as possible by means of a series of examples and the manifold difficulties met with as well as the methods by which we seek to overcome these difficulties.

2. Let us commence our consideration with the examination of the magnitude δ, to which various methods of comprehensive presentation of stochastical connexion refer if both the variables can assume only two different values each (cf. Chap. IV, §§ 2 and 7). If we define the *a priori* magnitude δ as $\delta = p_{1|1}p_{2|2} - p_{1|2}p_{2|1}$, and if we substitute for the *a priori* probabilities the corresponding empirical frequencies, then we arrive at a function of empirical values which we have denoted by δ' (cf. Chap. V, § 7):

$$\delta' = p'_{1|1}p'_{2|2} - p'_{1|2}p'_{2|1} = \frac{1}{N^2}[n_{1|1}n_{2|2} - n_{1|2}n_{2|1}].$$

Hence it is of importance to ascertain how far the value of δ' is suitable to serve as a presumptive value of δ, and whether there are better methods of estimation.

Because, as it is known,

$$\mathsf{E}n_{1|1}n_{2|2} = N(N-1)p_{1|1}p_{2|2}, \quad \mathsf{E}n_{1|2}n_{2|1} = N(N-1)p_{1|2}p_{2|1}$$

so is

$$\mathsf{E}\delta' = \frac{N-1}{N}\delta.$$

Consequently the value of δ would be systematically underrated should δ' be considered as a presumptive value of δ. The magnitude of the systematic error is of order $\frac{1}{N}$; hence it is not of great consequence in a fairly great number of trials. In this case it can be easily eliminated: it is only necessary to put the presumptive value of δ not equal to

δ', but equal to $\frac{N}{N-1}\delta'$; for the mathematical expectation of $\frac{N}{N-1}\delta'$ is equal to the *a priori* magnitude δ in any finite number of trials.

If we wish to gain an idea of the certainty with which the true value of δ can be estimated on the basis of empirical values $\frac{N}{N-1}\delta'$, we must ascertain the standard error of $\frac{N}{N-1}\delta'$. I am not showing the computation in detail as it is of no particular interest, but give only the final results:

$$\sigma^2_{\frac{N}{N-1}\delta'} = \mathrm{E}\left[\frac{N}{N-1}\delta' - \delta\right]^2 =$$

$$= \frac{1}{N-1}\{[p_{1|1}p_{2|2}(p_{1|1}+p_{2|2}) + p_{1|2}p_{2|1}(p_{1|2}+p_{2|1}) - 4\delta^2] +$$

$$+ \frac{1}{N}[p_{1|2} + p_{2|1} - p_{1|1} - p_{2|2} + 6\delta]\delta\}.$$

If $\delta = 0$, the mathematical expectation of $\frac{N}{N-1}\delta'$ likewise equals zero, and the standard error of $\frac{N}{N-1}\delta'$ becomes equal to $\frac{1}{N-1}p_{1|}p_{2|}p_{|1}p_{|2}$, as for $\delta = 0$ the associations consist of $p_{1|1} = p_{1|}p_{|1}$, $p_{1|2} = p_{1|}p_{|2}$.

Let us assume on the other hand that $\frac{N}{N-1}\delta'$ equals zero. We cannot conclude from this without any further consideration that δ likewise equals zero and the variables are mutually independent. It might happen that δ' equals zero by chance, although δ is different from zero. Nor can it be inferred without further consideration from the fact that $\frac{N}{N-1}\delta'$ is different from zero that δ is different from zero and the two variables are not independent. Here, also, the hand of chance can play its part. The plausibility of the conclusion depends in both cases on the magnitude of the standard error of $\frac{N}{N-1}\delta'$ and increases, as can be seen

from the above-mentioned formula for the standard error of $\frac{N}{N-1}\delta'$ with an increasing number of trials, proportionally to the square root of N.

If we wished to ascertain with greater exactness the frequency-distribution of the values of $\frac{N}{N-1}\delta'$ we should have to consider the mathematical expectations of higher powers of $\frac{N}{N-1}\delta'$. The computations are somewhat laborious, but do not offer any particular difficulties, and can be carried out in the same manner as the computation of the mathematical expectation and of the variance of $\frac{N}{N-1}\delta'$. The distribution of $\frac{N}{N-1}\delta'$-values is asymmetrical, but it approaches the Gauss-Laplace's form with the increase of the number of trials.

3. We meet quite different difficulties when we have to estimate the value of a quotient such as appears in a correlation coefficient, on the basis of empirical material in the case of variables which can assume only two possible values each (cf. Chap. IV, § 7):

$$r_{1|1} = \frac{\delta}{\sqrt{p_{1|}p_{2|}p_{|1}p_{|2}}},$$

viz. if by the usual method of the substitution of statistical frequencies for the *a priori* probabilities (Chap. V, § 7) we form the expression

$$r'_{1|1} = \frac{\delta'}{\sqrt{p'_{1|}p'_{2|}p'_{|1}p'_{|2}}},$$

then the mathematical expectation of $r'_{1|1}$ cannot be exactly calculated, at least not in the present state of our knowledge, because it is a quotient (cf. above, § 2, 2). We depend on approximations with which we will make a closer acquaintance in the general problem of the estimation of the value of *a priori* correlation coefficients (cf. below, § 4, 3, A, and § 4, 5).

§ 4

1. The difficulties of the calculation of the mathematical expectation of quotients play such an important part in the theory of methods of investigation of stochastically connected chance variables that we shall consider the problem more closely.

The mathematical expectation of a quotient has so far been precisely determined in a few exceptional cases only. The first case seems to be the calculation of the mathematical expectation of Lexis' divergence quotient. Some cases can also be found in the field of the theory of correlation.

A. According to the definition the value of $\mathsf{E}\frac{x}{y}$ equals $\sum_i \sum_j p_{i|j} \frac{x_i}{y_j}$. Accordingly the next thing would be to insert the value of the probability $p_{i|j}$ and to carry out the double summation. As a rule, this is not feasible, as the summations are intractable. Let us, for instance, calculate the mathematical expectation of $\frac{n_{i|j}}{n_{i|}}$. The probability that $n_{i|}$ takes the value h equals $\binom{N}{h} p_{i|}^h (1 - p_{i|})^{N-h}$. The conditional mathematical expectation of $n_{i|j}$, under the assumption that $n_{i|}$ equals h, comes to $h\frac{p_{i|j}}{p_{i|}}$, as the h-cases are distributed among sub-groups $n_{i|1}, n_{i|2} \ldots, n_{i|j} \ldots, n_{i|l}$ with probabilities $p_{i|1}, p_{i|2} \ldots, p_{i|l}$, reduces to the sum $p_{i|}$. Accordingly, the mathematical expectation of $\frac{n_{i|j}}{n_{i|}}$ equals:

$$\mathsf{E}\frac{n_{i|j}}{n_{i|}} = \sum_h \binom{N}{h} p_{i|}^h (1 - p_{i|})^{N-h} \frac{1}{h} h \frac{p_{i|j}}{p_{i|}} =$$

$$= \frac{p_{i|j}}{p_{i|}} \sum_h \binom{N}{h} p_{i|}^h (1 - p_{i|})^{N-h} = \frac{p_{i|j}}{p_{i|}}.$$

In this case the computation does not offer any difficulties, because h in the numerator and denominator of the values

to be summed up cancels out. But if we wanted to compute the mathematical expectation of $\frac{n_{i|j}^2}{n_{i|}^2}$ in the same manner we should have, for the conditional mathematical expectation of $n_{i|j}^2$ to insert, under the assumption that $n_{i|}$ equals h, $h^2\frac{p_{i|j}^2}{p_{i|}^2} + h\frac{p_{i|j}}{p_{i|}}\left(1 - \frac{p_{i|j}}{p_{i|}}\right)$ and our computation would become

$$\mathsf{E}\frac{n_{i|j}^2}{n_{i|}^2} = \sum_h \binom{N}{h} p_{i|}^h (1 - p_{i|})^{N-h} \frac{1}{h^2}\left[h^2\frac{p_{i|j}^2}{p_{i|}^2} + h\frac{p_{i|j}}{p_{i|}}\left(1 - \frac{p_{i|j}}{p_{i|}}\right)\right] =$$

$$= \frac{p_{i|j}^2}{p_{i|}^2} + \frac{p_{i|j}}{p_{i|}}\left(1 - \frac{p_{i|j}}{p_{i|}}\right)\sum_h \binom{N}{h} p_{i|}^h (1 - p_{i|})^{N-h}\frac{1}{h}.$$

Hence, we should arrive at a dead end, as the summation $\sum_h \binom{N}{h} p_{i|}^h (1 - p_{i|})^{N-h}\frac{1}{h}$ cannot be carried out precisely.

B. Sometimes the goal which can be reached directly can be attained indirectly. It can, for instance, be shown as follows, that the mathematical expectation of the empirical correlation coefficient (Chap. V, § 2, 2) in the case of mutual independence of variables is precisely zero and the standard error of the empirical correlation coefficient equals in this case precisely $\sqrt{\frac{1}{N-1}}$.

If we introduce, to facilitate the writing, the abbreviated notations

$$\sqrt{\left\{\sum_{f=1}^{N}[x^{[f]\prime} - x_0']^2\right\}} = \Sigma_1, \quad \sqrt{\left\{\sum_{f=1}^{N}[y^{[f]\prime} - y_0']^2\right\}} = \Sigma_2,$$

then the empirical correlation coefficient is defined by the formula

$$r'_{1|1} = \frac{\sum_{f=1}^{N}[x^{[f]\prime} - x_0'][y^{[f]\prime} - y_0']}{\Sigma_1\Sigma_2}.$$

Now, assume that the law of independence remains the

same at all trials and that the latter are mutually independent, then

$$\mathsf{E}\frac{x^{[f]\prime}}{\Sigma_1} = \mathsf{E}\frac{x^{[a]\prime}}{\Sigma_1} = \mathsf{E}\frac{x'_0}{\Sigma_1} \text{ and hence } \mathsf{E}\frac{x^{[f]\prime} - x'_0}{\Sigma_1} = 0.$$

Similarly we find $\mathsf{E}\dfrac{y^{[f]\prime} - y'_0}{\Sigma_2} = 0.$

When the variables X and Y are mutually independent

$$\mathsf{E}\frac{[x^{[f]\prime} - x'_0]\,[y^{[f]\prime} - y'_0]}{\Sigma_1\Sigma_2} = \left\{\mathsf{E}\frac{x^{[f]\prime} - x'_0}{\Sigma_1}\right\}\left\{\mathsf{E}\frac{y^{[f]\prime} - y'_0}{\Sigma_2}\right\}.$$

Hence

$$\mathsf{E}r'_{1|1} = \mathsf{E}\frac{\sum_{f=1}^{N}[x^{[f]\prime} - x'_0]\,[y^{[f]\prime} - y'_0]}{\Sigma_1\Sigma_2} =$$

$$= \sum_{f=1}^{N}\left\{\mathsf{E}\frac{[x^{[f]\prime} - x'_0]\,[y^{[f]\prime} - y'_0]}{\Sigma_1\Sigma_2}\right\} =$$

$$= \sum_{f=1}^{N}\left\{\left(\mathsf{E}\frac{x^{[f]\prime} - x'_0}{\Sigma_1}\right)\left(\mathsf{E}\frac{y^{[f]\prime} - y'_0}{\Sigma_2}\right)\right\} = 0.$$

Consequently, in the case of mutual independence of the variables X and Y the value of the mathematical expectation of $r'_{1|1}$ coincides with the true value of the *a priori* correlation coefficient $r_{1|1}$, which in the case of mutual independence of variables also equals zero (cf. Chap. IV, § 3, 1).

It must be borne in mind that our computations proceed from the assumption of the mutual independence of variables in the sense of our strong definition (cf. Chap. III, § 4, 3) and not from the assumption $r_{1|1} = 0$, which, as we know, is a necessary but not sufficient condition of mutual independence (cf. Chap. IV, § 3, 1 and § 3, 3). When $r_{1|1}$ equals 0, but the variables X and Y are not mutually independent, we must not assume that

$$\mathsf{E}\frac{[x^{[f]\prime} - x'_0]\,[y^{[f]\prime} - y'_0]}{\Sigma_1\Sigma_2} = \left\{\mathsf{E}\frac{x^{[f]\prime} - x'_0}{\Sigma_1}\right\}\cdot\left\{\mathsf{E}\frac{y^{[f]\prime} - y'_0}{\Sigma_2}\right\}.$$

The mathematical expectation of $r'_{1|1}$ can, as we shall see

later (cf. *infra*, § 4, 3, A), at $r_{1|1} = 0$ equal 0 as well as > 0 and < 0.

Similarly the standard error of $r'_{1|1}$ can be computed in the case of mutual independence of variables X and Y. Bearing in mind that among our assumptions

$$\mathsf{E}\frac{[x^{(f)\prime} - x'_0]\,[x^{(d)\prime} - x'_0]}{\Sigma_1^2} = -\frac{1}{N-1}\mathsf{E}\frac{[x^{(f)\prime} - x'_0]^2}{\Sigma_1^2}$$

and

$$\mathsf{E}\frac{[x^{(f)\prime} - x'_0]^2}{\Sigma_1^2} = \frac{1}{N},$$

we find easily that $\mathsf{E}[r'_{1|1}]^2 = \frac{1}{N-1}$. But as $\mathsf{E}r'_{1|1} = 0$

$$\sigma^2_{r'_{1|1}} = \mathsf{E}[r'_{1|1} - \mathsf{E}r'_{1|1}]^2 = \frac{1}{N-1}.$$

Both the results thus obtained, $\mathsf{E}r'_{1|1} = 0$, $\sigma^2_{r'_{1|1}} = \frac{1}{N-1}$, hold good for any laws of distribution of X and Y. On the other hand, the mathematical expectation of higher powers of $r'_{1|1}$ cannot be ascertained in a similar way, even in the case of mutual independence of the variables X and Y. For instance, in the computation of $\mathsf{E}[r'_{1|1}]^4$ we meet the mathematical expectations

$$\mathsf{E}\frac{\sum_{f=1}^{N}[x^{(f)\prime} - x'_0]^4}{\Sigma_1^4} \quad \text{and} \quad \mathsf{E}\frac{\sum_{f=1}^{N}[y^{(f)\prime} - y'_0]^4}{\Sigma_2^4},$$

which cannot be computed until the laws of distribution of X and Y have been determined.

2. As the final aim of the exact computation of $\mathsf{E}\frac{x}{y}$ cannot be achieved either directly or indirectly, nothing remains but to seek to determine approximately the mathematical expectation of a function of empirical values selected for the estimation of *a priori* magnitude. An obvious idea is to represent the sought-for mathematical expectation of U' as a sum of terms, arranged according to increased powers of $\frac{1}{N}$, assuming that these terms rapidly decrease

when N is sufficiently large. This fundamental idea can be put in different forms. The favourite method is the following due to the English School.

A. Let us compute the mathematical expectation of $\frac{n'^2_{i|j}}{n'^2_{i|}}$ or $\mathsf{E}\left[\frac{p'_{i|j}}{p'_{i|}}\right]^2$. If, for brevity, we introduce the notations $p'_{i|j} - p_{i|j} = dp'_{i|j}$, $p'_{i|} - p_{i|} = dp'_{i|}$, then $\left[\frac{p'_{i|j}}{p'_{i|}}\right]^2$ can be expanded according to increasing powers of $dp'_{i|j}$ and $dp'_{i|}$ as follows:

$$\left[\frac{p'_{i|j}}{p'_{i|}}\right]^2 = \frac{[p_{i|j} + dp'_{i|j}]^2}{[p_{i|} + dp'_{i|}]^2} = \frac{p^2_{i|j}}{p^2_{i|}}\left[1 + \frac{dp'_{i|j}}{p^2_{i|j}}\right]^2\left[1 + \frac{dp'_{i|}}{p_{i|}}\right]^{-2} =$$

$$= \frac{p^2_{i|j}}{p^2_{i|}}\left[1 + \frac{2dp'_{i|j}}{p_{i|j}} + \frac{(dp'_{i|j})^2}{p^2_{i|j}}\right]\left[1 - \frac{2dp'_{i|}}{p_{i|}} + \frac{3(dp'_{i|})^2}{p^2_{i|}} - \ldots\right] =$$

$$= \frac{p^2_{i|j}}{p^2_{i|}}\left[1 + \frac{2dp'_{i|j}}{p_{i|j}} - \frac{2dp'_{i|}}{p_{i|}} + \frac{(dp'_{i|j})^2}{p^2_{i|j}} - \frac{4dp'_{i|j}dp'_{i|}}{p_{i|j}p_{i|}} + \frac{3(dp'_{i|})^2}{p^2_{i|}} + \ldots\right].$$

Hence we obtain:

$$\mathsf{E}\left[\frac{p'_{i|j}}{p'_{i|}}\right]^2 = \frac{p^2_{i|j}}{p^2_{i|}}\left\{1 + \mathsf{E}\left[\frac{2dp'_{i|j}}{p_{i|j}} - \frac{2dp'_{i|}}{p_{i|}}\right] + \mathsf{E}\left[\frac{(dp'_{i|j})^2}{p^2_{i|j}} - \right.\right.$$

$$\left.\left. - \frac{4dp'_{i|j}dp'_{i|}}{p_{i|j}p_{i|}} + \frac{3(dp'_{i|})^2}{p^2_{i|}}\right] + \ldots\right\}.$$

Now as $\mathsf{E}\{[dp'_{i|j}]^h[dp'_{i|}]^{2s-1-h}\}$ and $\mathsf{E}\{[dp'_{i|j}]^h[dp'_{i|}]^{2s-h}\}$ contain no terms of order in $\frac{1}{N}$ lower than $\left(\frac{1}{N}\right)^s$, we can, proceeding from the above expansion, compute the sought-for mathematical expectation of $\left[\frac{p'_{i|j}}{p'_{i|}}\right]^2$ to the desired approximation of order $\frac{1}{N}$. If the desired approximation does commence with terms of order $\frac{1}{N}$ the series may be separated as has been done above. If we wish to be correct to terms of order $\left(\frac{1}{N}\right)^2$, then terms of the

3rd and 4th order in $dp'_{i|j}$ and $dp'_{i|}$ must be retained. In this way we obtain for $\mathsf{E}\left[\frac{p'_{i|j}}{p'_{i|}}\right]^2$

$$\mathsf{E}\frac{n^2_{i|j}}{n^2_{i|}} = \mathsf{E}\left[\frac{p'_{i|j}}{p'_{i|}}\right]^2 = \frac{p^2_{i|j}}{p^2_{i|}}\left\{1 + \frac{1}{N}\left[\frac{p_{i|j}(1-p_{i|j})}{p^2_{i|j}} - \frac{4p_{i|j}(1-p_{i|})}{p_{i|j}p_{i|}} + \right.\right.$$

$$\left.\left. + \frac{3p_{i|}(1-p_{i|})}{p^2_{i|}}\right] + \ldots\right\} = \frac{p^2_{i|j}}{p^2_{i|}}\left\{1 + \frac{p_{i|}-p_{i|j}}{Np_{i|}p_{i|j}} + \right.$$

$$\left. + \frac{(p_{i|}-p_{i|j})(1-p_{i|})}{N^2p^2_{i|}p_{i|j}} + \ldots\right\}.$$

B. Similarly from

$$\frac{n^2_{i|j}}{n_{i|}n_{|}} = \frac{p^2_{i|j}}{p_{i|}p_{|j}} \frac{\left[1 + \frac{dp'_{i|j}}{p_{i|j}}\right]^2}{\left[1 + \frac{dp'_{i|}}{p_{i|}}\right]\left[1 + \frac{dp'_{|j}}{p_{|j}}\right]}$$

can be obtained

$$\mathsf{E}\frac{n^2_{i|j}}{n_{i|}n_{|j}} = \frac{p^2_{i|j}}{p_{i|}p_{|j}}\left\{1 + \frac{[p_{i|}-p_{i|j}][p_{|j}-p_{i|j}]}{Np_{i|}p_{|j}p_{i|j}} + \right.$$

$$\left. + \frac{[p_{i|}-p_{i|j}][p_{|j}-p_{i|j}][2p_{i|j}-p_{i|}p_{|j}]}{N^2p^2_{i|}p^2_{|j}p_{i|j}} + \ldots\right\}.$$

C. The last formula permits a critical appreciation of the usual method of estimation of the value of *a priori* mean square contingency from empirical material. If we define the *a priori* mean square contingency as

$$\varphi^2 = \sum_i\sum_j \frac{[p_{i|j}-p_{i|}p_{|j}]^2}{p_{i|}p_{|j}} = \sum_i\sum_j \frac{p^2_{i|j}}{p_{i|}p_{|j}} - 1$$

and substitute frequencies p' for probabilities p in the empirical expression

$$[\varphi']^2 = \sum_i\sum_j \frac{[p'_{i|j}-p'_{i|}p'_{|j}]^2}{p'_{i|}p'_{|j}} = \sum_i\sum_j \frac{[n_{i|j}-\frac{1}{N}n_{i|}n_{|j}]^2}{n_{i|}n_{|j}} =$$

$$= \sum_i\sum_j \frac{n^2_{i|j}}{n_{i|}n_{|j}} - 1,$$

then the value of $[\varphi']^2$ holds as a presumptive value of φ^2.

The mathematical expectation of $[\varphi']^2$ is obtained by inserting the above value of $\mathsf{E}\frac{n_{i|j}^2}{n_{i|}n_{|j}}$:

$$\mathsf{E}[\varphi']^2 = \varphi^2 + \frac{1}{N}\left\{\sum_i\sum_j \frac{p_{i|j}[p_{i|} - p_{i|j}][p_{|j} - p_{i|j}]}{p_{i|}^2 p_{|j}^2}\right\} +$$

$$+ \frac{1}{N^2}\left\{\sum_i\sum_j \frac{p_{i|j}(p_{i|} - p_{i|j})(p_{|j} - p_{i|j})(2p_{i|j} - p_{i|}p_{|j})}{p_{i|}^3 p_{|j}^3}\right\} + \ldots$$

We satisfy ourselves that the mathematical expectation of $[\varphi']^2$ is larger than φ^2, and we consequently systematically overestimate the value of the *a priori* mean square contingency, if we take the value $[\varphi']^2$ as its presumptive value. We are not, however, able at present to eliminate this systematic error as we were in the case of δ: it cannot be seen from our expansion in series how $[\varphi']^2$ can be modified to give a function of empirical values, the mathematical expectation of which precisely equals the *a priori* value of φ^2 in any finite number of trials.

When the variables X and Y are mutually independent $\varphi^2 = 0$, we obtain for the mathematical expectation of $[\varphi']^2$ the value

$$\mathsf{E}[\varphi']^2 = \frac{(k-1)(l-1)}{N} + \frac{(k-1)(l-1)}{N^2} + \ldots$$

It is most probable that the mathematical expectation of $[\varphi']^2$ when X and Y are mutually independent, is precisely $\frac{(k-1)(l-1)}{N-1}$, but I have not yet succeeded in proving this. If it were really so, we could obtain an empirical value at least for the case of mutual independence of X and Y by the subtraction of $\frac{(k-1)(l-1)}{N-1}$ from the value of $[\varphi']^2$, the mathematical expectation of which would under this assumption precisely equal the value of φ^2.

Similarly we can compute the variance of $[\varphi']^2$. In

the general case of any law of dependence we find:

$$\sigma^2_{[\varphi']^2} = \mathrm{E}\{[\varphi']^2 - \mathrm{E}[\varphi']^2\}^2 = \frac{1}{N}\Big\{4\sum_i\sum_j \frac{p^3_{i|j}}{p^2_{i|}p^2_{|j}} -$$

$$- 3\sum_i\Big[\frac{1}{p^3_{i|}}\Big(\sum_j \frac{p^2_{i|j}}{p_{|j}}\Big)^2\Big] - 3\sum_j\Big[\frac{1}{p^3_{|j}}\Big(\sum_i \frac{p^2_{i|j}}{p_{i|}}\Big)^2\Big] +$$

$$+ 2\sum_i\sum_j\Big[\frac{p_{i|j}}{p^2_{i|}p^2_{|j}}\Big(\sum_{f=1}^{k}\frac{p^2_{f|j}}{p_{f|}}\Big)\Big(\sum_{g=1}^{l}\frac{p^2_{i|g}}{p_{|g}}\Big)\Big]\Big\} + \ldots$$

In the case of mutual independence of the variables X and Y the term of order $\frac{1}{N}$ in $\sigma^2_{[\varphi']_2}$ disappears. Hence the variance of $[\varphi']^2$ is in the case of mutual independence of the variables of order $\left(\frac{1}{N}\right)^2$ and the standard error of $[\varphi']^2$ of order $\frac{1}{N}$.

D. Similarly we can compute the mathematical expectations of all those functions of empirical values which are formed by the insertion of frequencies instead of probabilities in the formulae which define the *a priori* magnitudes. The computations are mostly very laborious—particularly when we are not satisfied with the computation of the term of the order $\frac{1}{N}$, but they do not cause any difficulties even when greater precision is aimed at. Only we must not overlook the fact that the mathematical expectation of an odd function of differences $dp'_{i|}$, &c., always contains terms of the same order of magnitude $\frac{1}{N}$, as the mathematical expectation of the next highest even function, so that in any attempt at greater precision special care must be paid to the expansion of the series of powers of differences $dp'_{i|}$, &c., to include not only terms next in order but also the even terms immediately following. English statisticians have sometimes fallen into serious error through the non-observance of these rules.

3. Laborious calculations can be simplified in many cases

by introducing differences between certain functions of frequencies and the mathematical expectations of these functions instead of differences of frequencies and of probabilities.

A. For instance, we put

$$\mu''_{r|s} = \sum_{i|j} p'_{i|j}[x_i - m_{1|0}]^r[y_j - m_{0|1}]^s,\ d\mu''_{r|s} = \mu''_{r|s} - \mu_{r|s},$$

whereas, since $\mu_{1|0} = \mu_{0|1} = 0$,

$$m'_{1|1} - m'_{1|0}m'_{0|1} = \mu''_{1|1} - \mu''_{1|0}\mu''_{0|1} = \mu_{1|1} + d\mu''_{1|1} - d\mu''_{1|0}\, d\mu''_{0|1}$$

$$m'_{2|0} - m'^2_{1|0} = \mu'_{2|0} - \mu''^2_{1|0} = \mu_{2|0} + d\mu''_{2|0} - (d\mu''_{1|0})^2$$

$$m'_{0|2} - m'^2_{0|1} = \mu''_{0|2} - \mu''^2_{0|1} = \mu_{0|2} + d\mu''_{0|2} - (d\mu''_{0|1})^2.$$

The empirical expression usually employed for the estimation of the value of *a priori* correlation coefficient can be then represented in the following form :

$$r'_{1|1} = \frac{\mu''_{1|1} - \mu''_{1|0}\mu''_{0|1}}{\sqrt{[\mu''_{2|0} - \mu''^2_{0|1}]\,[\mu''_{0|2} - \mu''^2_{0|1}]}} =$$

$$= r_{1|1}\frac{\left[1 + \frac{d\mu''_{1|1}}{\mu_{1|1}} - \frac{d\mu''_{1|0}\, d\mu''_{0|1}}{\mu_{1|1}}\right]}{\left[1 + \frac{d\mu''_{2|0}}{\mu_{2|0}} - \frac{(d\mu''_{1|0})^2}{\mu_{2|0}}\right]^{\frac{1}{2}}\left[1 + \frac{d\mu''_{0|2}}{\mu_{0|2}} - \frac{(d\mu''_{0|2})^2}{\mu_{0|2}}\right]^{\frac{1}{2}}}.$$

If we expand in powers of differences $d\mu''_{1|1}$, $d\mu''_{1|0}$, &c., and bear in mind that the mathematical expectations of $[d\mu''_{1|1}]^{2s-1}$ and of $[d\mu''_{1|1}]^{2s}$ comprise no terms of lower order in $\frac{1}{N}$ than $\left(\frac{1}{N}\right)^s$, &c., then we obtain the value of $\mathsf{E}r'_{1|1}$ with the desired precision in $\frac{1}{N}$ by interrupting the series at the corresponding terms of even order and inserting the mathematical expectations of $[d\mu''_{1|1}]^h$, $[d\mu''_{1|0}d\mu''_{0|1}]^h$, &c. ; the computation does not involve any fundamental difficulties. In this manner we obtain

$$\mathsf{E}r'_{1|1} = r_{1|1} + \frac{1}{N}\left\{\frac{1}{4}r_{2|2}r_{1|1} + \frac{3}{8}r_{1|1}[r_{4|0} + r_{0|4}] - \frac{1}{2}[r_{3|1} + r_{1|3}]\right\} +$$

$$+ \frac{1}{N^2}\left\{\frac{1}{4}r_{3|3} - \frac{5}{16}r_{1|1}[r_{6|0} + r_{0|6}] - \frac{3}{16}r_{1|1}[r_{4|2} + r_{2|4}] +\right.$$

$$+\frac{3}{8}[r_{5|1}+r_{1|5}]-\frac{1}{4}r_{2|2}r_{1|1}+\frac{1}{2}[r_{3|1}+r_{1|3}]+$$
$$+\frac{15}{32}r_{1|1}r_{2|2}[r_{4|0}+r_{0|4}]-\frac{3}{8}r_{1|1}[r_{4|0}+r_{0|4}]+$$
$$+\frac{9}{64}r_{1|1}r_{4|0}r_{0|4}+\frac{9}{32}r_{1|1}r_{2|2}^2-\frac{15}{16}[r_{3|1}r_{4|0}+r_{1|3}r_{0|4}]-$$
$$-\frac{3}{16}[r_{3|1}r_{0|4}+r_{1|3}r_{4|0}]-\frac{3}{8}r_{2|2}[r_{3|1}+r_{1|3}]+$$
$$+\frac{105}{128}r_{1|1}[r_{4|0}^2+r_{0|4}^2]-\frac{1}{4}r_{3|0}r_{0|3}+\frac{15}{8}r_{1|1}[r_{3|0}^2+r_{0|3}^2]-$$
$$-\frac{9}{4}[r_{2|1}r_{3|0}+r_{1|2}r_{0|3}]-\frac{5}{4}r_{2|1}r_{1|2}+\frac{3}{8}r_{1|1}[r_{2|1}^2+r_{1|2}^2]+$$
$$+\frac{3}{4}r_{1|1}[r_{2|1}r_{0|3}+r_{1|2}r_{3|0}]-\frac{1}{2}r_{1|1}+\frac{1}{2}r_{1|1}^3\}+\ldots$$

This formula holds good for any law of dependence. Hence with linear regression of X on Y and of Y on X we find, since with these assumptions

$$r_{\lambda|1}=r_{1|1}r_{\lambda+1|0} \quad \text{and} \quad r_{1|\lambda}=r_{1|1}r_{0|\lambda+1}:$$

$$\mathsf{E}r'_{1|1}=r_{1|1}+\frac{1}{N}r_{1|1}\left\{\frac{1}{4}r_{2|2}-\frac{1}{8}[r_{4|0}+r_{0|4}]\right\}+\frac{1}{N^2}\left\{\frac{1}{4}r_{3|3}+\frac{1}{16}r_{1|1}\right.$$
$$[r_{6|0}+r_{0|6}]-\frac{3}{16}r_{1|1}[r_{4|2}+r_{2|4}]-\frac{1}{4}r_{1|1}r_{2|2}+\frac{9}{32}r_{1|1}r_{2|2}^2+$$
$$+r_{1|1}[r_{4|0}+r_{0|4}]\left[\frac{1}{8}+\frac{3}{32}r_{2|2}\right]-\frac{15}{128}r_{1|1}[r_{4|0}+r_{0|4}]^2-$$
$$\left.-[1-r_{1|1}^2]\left[\frac{1}{2}r_{1|1}+\frac{1}{4}r_{3|0}r_{0|3}+\frac{3}{8}r_{1|1}(r_{3|0}^2+r_{0|3}^2)\right]\right\}+\ldots$$

When $r=0$ we find for any law of dependence

$$\mathsf{E}r'_{1|1}=-\frac{1}{2N}[r_{3|1}+r_{1|3}]+\frac{1}{N^2}\left\{\frac{1}{4}r_{3|3}+\frac{3}{8}[r_{5|1}+r_{1|5}]+\right.$$
$$+\frac{1}{2}[r_{3|1}+r_{1|3}]-\frac{15}{16}[r_{3|1}r_{4|0}+r_{1|3}r_{0|4}]-\frac{3}{16}[r_{3|1}r_{0|4}+$$
$$+r_{1|3}r_{4|0}]-\frac{3}{8}r_{2|2}[r_{3|1}+r_{1|3}]-\frac{1}{4}r_{3|0}r_{0|3}-\frac{9}{4}[r_{2|1}r_{3|0}+$$
$$\left.+r_{1|2}r_{0|3}]-\frac{5}{4}r_{2|1}r_{1|2}\right\}+\ldots$$

Hence $\mathsf{E}r'_{1|1}$ is different from zero and can be positive as well as negative, according to the sign of $[r_{3|1}+r_{1|3}]$.

In the case of mutual independence of the variables X and Y we obtain, within the limits of our approximation, $\mathsf{E}r' = 0$, since $r_{\varrho|f} = r_{\varrho|0}r_{0|f}$ and $r_{1|0} = r_{0|1} = 0$.

In the case of ' normal correlation ' by substituting their expressions for the higher r-parameters we find by means of $r_{1|1}$ (cf. above, Chap. IV, § 5, 2) :

$$\mathsf{E}r'_{1|1} = r\left\{1 - \frac{1 - r^2_{1|1}}{2N} - \frac{3[1 - r^2_{1|1}]\,[1 + 3r^2_{1|1}]}{8N^2} \cdots\right\}.$$

Hence, the value of the coefficient of correlation in the case of normal correlation is systematically underrated if the presumptive value of $r_{1|1}$ is put equal to $r'_{1|1}$.

The general formula for $\mathsf{E}r'_{1|1}$, however, shows that the value of $r_{1|1}$ is not always underrated, when its presumptive value is put equal to $r'_{1|1}$; in the case of non-normal correlation it can also be over-estimated. An elimination of systematic errors by a modification of expression $r'_{1|1}$ which holds as a presumptive value, can no more be achieved on the basis of the results obtained up to the present than in the case of the mean square contingency. Still, the position in the case of $r'_{1|1}$ is more favourable since, with mutual independence of the variables, the value of $\mathsf{E}r'_{1|1}$ coincides with the true value of $r_{1|1}$, which then equals zero, as we have satisfied ourselves (cf. above, § 4, 1, B).

In a similar way the variance of $r'_{1|1}$ can be ascertained. We obtain to the first approximation,

$$\sigma^2_{r'_{1|1}} = \mathsf{E}\{r'_{1|1} - \mathsf{E}r'_{1|1}\}^2 =$$

$$= \frac{1}{N}\left\{r_{2|2}\left[1 + \frac{1}{2}r^2_{1|1}\right] - r_{1|1}[r_{3|1} + r_{1|3}] + \frac{1}{4}r^2_{1|1}[r_{4|0} + r_{0|4}]\right\} + \cdots$$

The same value is obtained to the first approximation for $\mathsf{E}\{r'_{1|1} - r_{1|1}\}^2$.

With normal correlation we find to the first approximation

$$\mathsf{E}\{r'_{1|1} - \mathsf{E}r'_{1|1}\}^2 = \mathsf{E}\{r'_{1|1} - r_{1|1}\}^2 = \frac{[1 - r^2_{1|1}]^2}{N} + \cdots$$

When the regression of Y on X is linear

$$\sigma^2_{r'_{1|1}} = \frac{1}{N}\left\{r_{2|2}\left[1 + \frac{1}{2}r^2_{1|1}\right] - \frac{3}{4}r^2_{1|1}r_{4|0} - r_{1|1}r_{1|3} + \frac{1}{4}r^2_{1|1}r_{0|4}\right\} + \cdots$$

If both the regressions of Y on X and that of X on Y are linear,

$$\sigma^2_{r'_{1|1}} = \frac{1}{N}\left\{r_{2|2}\left[1 + \frac{1}{2}r^2_{1|1}\right] - \frac{3}{4}r^2_{1|1}[r_{4|0} + r_{0|4}]\right\} + \ldots$$

With $r_{1|1} = 0$ we arrive at

$$\sigma^2_{r'_{1|1}} = \mathsf{E}[r'_{1|1}]^2 = \frac{r_{2|2}}{N} + \ldots$$

With mutual independence of the variables $r_{2|2} = r_{2|0}r_{0|2} = 1$, and consequently the first approximation is $\sigma^2_{r'_{1|1}} = \frac{1}{N}$. The precise value of the variance is, in this case, as we have seen above (cf. above, § 4, 1, B), $\frac{1}{N-1}$.

B. The problem of determination of presumptive values of coefficients of regression equations can be solved similarly. We should like to confine ourselves here to the problem of determining approximately the equation of the straight line of the best possible fit to the true regression line. As we have seen (cf. Chap. IV, § 3, 4) the *a priori* equation has the form

$$M^{(i)}_{|1} = \frac{m_{2|0}m_{0|1} - m_{1|1}m_{1|0}}{m_{2|0} - m^2_{1|0}} + \frac{m_{1|1} - m_{1|0}m_{0|1}}{m_{2|0} - m^2_{1|0}}x_i = A_{|0} + A_{|1}x_i.$$

If we determine, by means of the Method of Least Squares, the equation of a straight line which best represents the relation between the empirical value of the conditional mathematical expectation of Y and the corresponding X values, we obtain (Chap. V, § 3)

$$M^{(i)'}_{|1} = \frac{m'_{2|0}m'_{0|1} - m'_{1|1}m'_{1|0}}{m'_{2|0} - m'^2_{1|0}} + \frac{m'_{1|1} - m'_{1|0}m'_{0|1}}{m'_{2|0} - m'^2_{1|0}}x_i = A'_{|0} + A'_{|1}x_i,$$

where, as before, $m'_{f|g} = \sum_i \sum_j p'_{i|j}x'^f_i y'^g_j$ is inserted.

How far may this straight line hold good as a presumptive representation of the corresponding *a priori* straight line? Or in other words: how far have the values of A'_0 and A'_1 to be considered as presumptive values of corresponding coefficients in the *a priori* equation?

I am not giving the expansion of the series, as the computations take exactly the same course as in the case of correlation coefficients and do not contain anything particularly instructive, but merely show the final results. Up to the term of order $\frac{1}{N}$ we obtain approximately

$$\mathsf{E}A'_{|1} = A_{|1} + \frac{1}{N}\sqrt{\frac{\mu_{0|2}}{\mu_{2|0}}}[r_{4|0}r_{1|1} - r_{3|1}] + \ldots$$

$$\mathsf{E}A'_{|0} = A_{|0} + \frac{1}{N}\sqrt{\frac{\mu_{0|2}}{\mu_{2|0}}}\{[r_{3|0}r_{1|1} - r_{2|1}] - m_{1|0}[r_{4|0}r_{1|1} - r_{3|1}]\} + \ldots$$

These relations hold good for any shape of the true regression line. If the true regression of Y on X is linear, then $r_{h|1} = r_{1|1}r_{h+1|0}$ and consequently, within the limits of our approximation, $\mathsf{E}A'_{|1} = A_{|1} = a_{|1}$ and $\mathsf{E}A'_{|0} = A_{|0} = a_{|0}$. It may be supposed that for linear regression of Y on X the relations $\mathsf{E}A'_{|1} = a_{|1}$ and $\mathsf{E}A'_{|0} = a_{|0}$ are precise; I have, however, failed to prove this.

For the variances of $A'_{|1}$ and of $A'_{|0}$ we obtain similarly the values:

$$\sigma^2_{A'_{|1}} = \frac{1}{N}\frac{\mu_{0|2}}{\mu_{2|0}}[r_{2|2} + r_{4|0}r^2_{1|1} - 2r_{3|1}r_{1|1}] + \ldots$$

$$\sigma^2_{A'_{|0}} = \frac{1}{N}\frac{\mu_{0|2}}{\mu_{2|0}}\{\mu_{2|0}[1 - r^2_{1|1}] - 2m_{1|0}\sqrt{\mu_{2|0}}[r_{1|2} - 2r_{2|1}r_{1|1} + r_{3|0}r^2_{1|1}] + m^2_{1|0}[r_{2|2} + r_{4|0}r^2_{1|1} - 2r_{3|1}r_{1|1}]\} + \ldots$$

These relations hold good quite generally for any law of dependence. Hence with linear regression of Y on X we obtain:

$$\sigma^2_{A'_{|1}} = \frac{1}{N}\frac{\mu_{0|2}}{\mu_{2|0}}[r_{2|2} - r^2_{1|1}r_{4|0}] + \ldots$$

$$\sigma^2_{A'_{|0}} = \frac{1}{N}\frac{\mu_{0|2}}{\mu_{2|0}}\{\mu_{2|0}[1 - r^2_{1|1}] - 2m_{1|0}\sqrt{\mu_{2|0}}[r_{1|2} - r^2_{1|1}r_{3\,0}] + m^2_{1|0}[r_{2|2} - r^2_{1|1}r_{4|0}]\} + \ldots$$

If the regression of Y on X is linear and also homoscedastic, we obtain

$$\sigma^2_{A'|1} = \frac{1}{N}\frac{\mu_{0|2}}{\mu_{2|0}}[1 - r^2_{1|1}] + \ldots$$

$$\sigma^2_{A'|0} = \frac{1}{N}\frac{\mu_{0|2}}{\mu_{2|0}}[1 - r^2_{1|1}][\mu_{2|0} + m^2_{1|0}] + \ldots$$

or to the first approximation

$$\sigma^2_{A'|0} = [m^2_{1|0} + \mu_{2|0}]\sigma^2_{A'|1} = m_{2|0}\sigma^2_{A'|1}.$$

The last formulae hold good also in the case of a normal correlation as the regression of Y on X is then linear and homoscedastic.

C. Finally, by means of similar expansions in series we can ascertain the mathematical expectation and the standard error of $[\eta'_{y|x}]^2$ if we form as an empirical counterpart of the *a priori* correlation ratio of Y on X (cf. Chap. IV, § 4, 2)

$$\eta^2_{y|x} = \frac{\sum_i p_{i|}[m^{(i)}_{|1} - m_{0|1}]^2}{\sum_j p_{|j}[y_j - m_{0|1}]^2}$$

the empirical correlation ratio (cf. Chap. V, § 4)

$$[\eta'_{y|x}]^2 = \frac{\sum_i p'_{i|}[m^{(i)\prime}_{|1} - m'_{0|1}]^2}{\sum_j p'_{|j}[y_j - m'_{0|1}]^2}.$$

In the general case of any law of dependence we obtain

$$\mathrm{E}[\eta'_{y|x}]^2 = \eta^2_{y|x} + \frac{1}{N}\Big\{[r_{0|4} + 1]\eta^2_{y|x} - 1 + \frac{1}{\mu_{0|2}}\sum_i \mu^{(i)}_{|2} -$$

$$- \frac{1}{\mu^2_{0|2}}\Big[\sum_i p_{i|}(m^{(i)}_{|1} - m_{0|1})^4 + 5\sum_i p_{i|}(m^{(i)}_{|1} - m_{0|1})^2\mu^{(i)}_{|2} +$$

$$+ 2\sum_i p_{i|}(m^{(i)}_{|1} - m_{0|1})\mu^{(i)}_{|3}\Big]\Big\} + \ldots$$

$$\sigma^2_{[\eta'_{y|x}]^2} = \frac{1}{N}\{r_{0|4}\eta^4_{y|x} + \frac{1}{\mu^2_{0|2}}[(1 - 2\eta^2_{y|x})\sum_i p_i(m^{(i)}_{|1} - m_{0|1})^4 +$$

$$+ 2(2 - 5\eta^2_{y|x})\sum_i p_i(m^{(i)}_{|1} - m_{0|1})^2\mu^{(i)}_{|2} - 4\eta^2_{y|x}\sum_i p_{i|}(m^{(i)}_{|1} -$$

$$- m_{0|1})\mu^{(i)}_{|3}]\} + \ldots$$

Hence, if the regression of Y on X is linear, we obtain

$$\mathsf{E}[\eta'_{y|x}]^2 = r^2_{1|1} + \frac{1}{N}\Big\{\frac{1}{\mu_{0|2}}\Sigma_i \mu^{(i)}_{|2} - 1 + r^2_{1|1} + r^2_{1|1} r_{0|4} + r^2_{1|1} r_{2|2} - 2r_{1|1} r_{1|3}\Big\} + \ldots$$

$$\sigma^2[\eta'_{y|x}]^2 = \frac{1}{N} r^2_{1|1}\{2[2 + r^2_{1|1}] r_{2|2} + r^2_{1|1} r_{0|4} - 3r^2_{1|1} r_{4|0} - 4r_{1|1} r_{1|3}\} + \ldots$$

If, in addition, the connexion of Y with X is homoscedastic,

$$\mathsf{E}[\eta'_{y|x}]^2 = r^2_{1|1} + \frac{1}{N}\{[k-1][1 - r^2_{1|1}] + r^2_{1|1} r_{0|4} + r^2_{1|1} r_{2|2} - 2r_{1|1} r_{1|3}\} + \ldots$$

If the variable Y is non-correlated with X, then $\eta^2_{y|x} = 0$, $m^{(i)}_{|1} = m_{0|1}$, and consequently

$$\mathsf{E}[\eta'_{y|x}]^2 = \frac{1}{N}\Big[\frac{1}{\mu_{0|2}}\Sigma_i \mu^{(i)}_{|2} - 1\Big] + \ldots$$

In the case of the variance $[\eta'_{y|x}]^2$ the term of order $\frac{1}{N}$ disappears in the expansion in powers of $\frac{1}{N}$. Hence the standard error of $[\eta'_{y|x}]^2$ in the case where Y is non-correlated with X is of order $\frac{1}{N}$.

If Y is uncorrelated with X and the regression of Y on X is homoscedastic,

$$\mathsf{E}[\eta'_{y|x}]^2 = \frac{1}{N}[k-1] + \ldots$$

For normal correlation $\eta^2_{y|x} = r^2_{1|1}$ and

$$\mathsf{E}[\eta'_{y|x}]^2 = r^2_{1|1} + \frac{1}{N}[1 - r^2_{1|1}][k - 1 - 2r^2_{1|1}] + \ldots$$

$$\sigma^2_{[\eta'_{y|x}]^2} = \frac{1}{N} 4r^2_{1|1}[1 - r^2_{1|1}]^2 + \ldots$$

D. We have seen (cf. Chap. IV, § 4, 3) that $\eta^2_{y|x} = r^2_{1|1}$ represents a necessary and sufficient condition of linearity of regression of Y on X. If we put in $\eta^2_{y|x} - r^2_{1|1} = \zeta^2_{y|x}$, the value of $\zeta^2_{y|x}$ can consequently serve as a criterion whether the regression of X on Y is linear; if $\zeta^2_{y|x}$ equals zero the

regression of Y on X is linear; if $\zeta^2_{y|x}$ is different from zero, the regression of Y on X cannot be linear.

The estimation of the value of $\zeta^2_{y|x}$ on the basis of empirical material proceeds from the value of difference $[\zeta'_{y|x}]^2$ between the corresponding empirical magnitudes:

$$[\zeta'_{y|x}]^2 = [\eta'_{y|x}]^2 - [r'_{1|1}]^2.$$

By means of similar expansions in series, as above, we obtain under the assumption of any law of dependence

$$\mathsf{E}[r'_{1|1}]^2 = r^2_{1|1} + \frac{1}{N}\{r_{2|2}[1 + r^2_{1|1}] + r^2_{1|1}[r_{4|0} + r_{0|4}] - 2r_{1|1}[r_{3|1} + r_{1|3}]\} + \ldots$$

$$\sigma^2_{[r'_{1|1}]^2} = \frac{1}{N}r^2_{1|1}\{2[2 + r^2_{1|1}]r_{2|2} + r^2_{1|1}[r_{4|0} + r_{0|4}] - 4r_{1|1}[r_{3|1} + r_{1|3}]\} + \ldots$$

If the regression of Y on X is linear we obtain

$$\mathsf{E}[r'_{1|1}]^2 = r^2_{1|1} + \frac{1}{N}\{r_{2|2}[1 + r^2_{1|1}] + r^2_{1|1}r_{0|4} - r^2_{1|1}r_{4|0} - 2r_{1|1}r_{1|3}\} + \ldots$$

$$\sigma^2_{[r'_{1|1}]^2} = \frac{1}{N}r^2_{1|1}\{2[2 + r^2_{1|1}]r_{2|2} + r^2_{1|1}r_{0|4} - 3r^2_{1|1}r_{4|0} - 4r_{1|1}r_{1|3}\} + \ldots$$

If both the regressions are linear we have

$$\mathsf{E}[r'_{1|1}]^2 = r^2_{1|1} + \frac{1}{N}\{r_{2|2}[1 + r^2_{1|1}] - r^2_{1|1}[r_{4|0} + r_{0|4}]\} + \ldots$$

$$\sigma^2_{[r'_{1|1}]^2} = \frac{1}{N}r^2_{1|1}\{2[2 + r^2_{1|1}]r_{2|2} - 3r^2_{1|1}[r_{4|0} + r_{0|4}]\} + \ldots$$

For $r_{1|1} = 0$ we find for any form of the lines of regression

$$\mathsf{E}[r'_{1|1}]^2 = \frac{1}{N}r_{2|2} + \ldots$$

In the case of the variance of $[r'_{1|1}]^2$ the term of order $\frac{1}{N}$ disappears in the expansion in powers of $\frac{1}{N}$ when $r_{1|1} = 0$. Hence the standard error of $[r'_{1|1}]^2$ in the case where $r_{1|1} = 0$, viz. where Y is uncorrelated with X, is of order $\frac{1}{N}$.

In the case of mutual independence of the variables $r_{2|2} = 1$ and from the expansion of $\mathsf{E}[r'_{1|1}]^2$ in powers of $\frac{1}{N}$ we obtain the first approximation

$$\mathsf{E}[r'_{1|1}]^2 = \frac{1}{N}.$$

In this case the precise value of $\mathsf{E}[r'_{1|1}]^2$ equals (cf. above, § 4, 1, B) $\frac{1}{N-1}$.

For normal correlation we find

$$\mathsf{E}[r'_{1|1}]^2 = r^2_{1|1} + \frac{1}{N}[1 - r^2_{1|1}][1 - 2r^2_{1|1}] + \ldots$$

$$\sigma^2_{[r'_{1|1}]^2} = \frac{1}{N}4r^2_{1|1}[1 - r^2_{1|1}]^2 + \ldots$$

Comparing the mathematical expectation and the variance of $[r'_{1|1}]^2$ with those of $[\eta'_{y|x}]^2$ (cf. *supra*, § 4, 3, C) we conclude that when $\eta^2_{y|x} = r^2_{1|1}$ (i.e. for linear regression of Y on X), the standard error of the estimate of the square of *a priori* correlation coefficient $r_{1|1}$ on the basis of chance value of $[r'_{1|1}]^2$ equals the standard error of the estimate of $\eta^2_{y|x}$ on the basis of chance value of $[\eta'_{y|x}]^2$, but that the systematic errors of the estimates are different. For normal correlation the difference due to systematic error of the estimate is to the first approximation:

$$\frac{1}{N}[k - 2][1 - r^2_{1|1}] + \ldots$$

The difference due to systematic error of the estimate of $\eta^2_{y|x} = r^2_{1|1}$ on the basis of chance values of $[\eta'_{y|x}]^2$ and of $[r'_{1|1}]^2$ coincides with the systematic error of the estimate of $\zeta^2_{y|x}$ on the basis of chance values of $[\zeta'_{y|x}]^2$. It follows from the above formulae that under the assumption that the regression of Y on X is linear and consequently

$$\zeta^2_{y|x} = \eta^2_{y|x} - r^2_{1|1} = 0,$$

$$\mathsf{E}[\zeta'_{y|x}]^2 - \zeta^2_{y|x} = \mathsf{E}[\eta'_{y|x}]^2 - \mathsf{E}[r'_{1|1}]^2 =$$

$$= \frac{1}{N}\left\{\frac{1}{\mu_{0|2}}\Sigma\mu^{(i)}_{|2} - [1 - r^2_{1|1}] - r_{2|2} + r^2_{1|1}r_{4|0}\right\} + \ldots$$

If the regression is also homoscedastic,

$$\mathsf{E}[\zeta'_{y|x}]^2 = \frac{1}{N}\{[k-1]\,[1-r_{1|1}^2] - r_{2|2} + r_{1|1}^2 r_{4|0}\} + \ldots$$

For normal correlation

$$\mathsf{E}[\zeta'_{y|x}]^2 = \frac{1}{N}[k-2]\,[1-r_{1|1}^2] + \ldots$$

For the variance of $[\zeta'_{y|x}]^2$ we obtain in the general case of any law of dependence, after rather laborious computations, the value

$$\begin{aligned}
\sigma^2_{[\zeta'_{y|x}]^2} = \frac{1}{N}\{&[\eta^2_{y|x} - r^2_{1|1}]^2 r_{0|4} + [\eta^2_{y|x} - r^2_{1|1}]2r_{1|1}[2r_{1|3} - r_{1|1}r_{2|2}] + \\
&+ 4r^2_{1|1}[r_{2|2} - r_{1|1}r_{3|1}] + r^4_{1|1}r_{4|0} + \\
&+ [1 - 2(\eta^2_{y|x} - r^2_{1|1})]\frac{1}{\mu^2_{0|2}}\Sigma_i p_{i|}[m^{(i)}_{|1} - m_{0|1}]^4 + \\
&+ [4 - 10(\eta^2_{y|x} - r^2_{1|1})]\frac{1}{\mu^2_{0|2}}\Sigma_i p_{i|}[m^{(i)}_{|1} - m_{0|1}]^2\mu^{(i)}_{|2} - \\
&- 4[\eta^2_{y|x} - r^2_{1|1}]\frac{1}{\mu^2_{0|2}}\Sigma_i p_{i|}[m^{(i)}_{|1} - m_{0|1}]\mu^{(i)}_{|3} + \\
&+ 2r^2_{1|1}\frac{1}{\mu_{2|0}\mu_{0|2}}\Sigma_i p_{i|}[m^{(i)}_{|1} - m_{0|1}]^2[x_i - m_{1|0}]^2 - \\
&- 4r_{1|1}\frac{1}{\sqrt{\mu_{2|0}\mu^3_{0|2}}}\Sigma_i p_{i|}[m^{(i)}_{|1} - m_{0|1}]^3[x_i - m_{1|0}] - \\
&- 8r_{1|1}\frac{1}{\sqrt{\mu_{2|0}\mu^3_{0|2}}}\Sigma_i p_{i|}[m^{(i)}_{|1} - m_{0|1}]\,[x_i - m_{1|0}]\mu^{(i)}_{|2}\} + \ldots
\end{aligned}$$

If the regression of Y on X is linear the term of order $\frac{1}{N}$ disappears in the expansion of the variance of $[\zeta'_{y|x}]^2$ in powers of $\frac{1}{N}$. Hence, the standard error of the difference $[\eta'_{y|x}]^2 - [r'_{1|1}]^2$ is for $\eta^2_{y|x} = r^2_{1|1}$, i.e. in the case of linear regression of Y on X, of order $\frac{1}{N}$.

E. The computations can be arranged much more easily, viz. much more comprehensively because we avoid having to retrace the differences of $dp'_{i|}$, $dp'_{i|j}$, &c., and consider the differences $dm'_{1|0}$, $d\mu''_{1|0}$, &c., instead. Still, they remain

quite heavy and they put the attention and the patience of the computer to the utmost test, although this performance does not involve any fundamental difficulties. The basic assumption that the mathematical expectations of higher powers of differences in question contain only correspondingly higher terms in $\frac{1}{N}$ is correct in all cases where the functions of empirical values used can be expanded in series in powers of differences $dp'_{i|}$, $dp'_{i|j}$, &c. Suppose that a function of empirical values of z' can be represented by $z' = c + d_1 + d_2 + d_3 + d_4 + \ldots$, where c denotes a constant, d_1 a sum of terms which contains differences $dp'_{i|}$, &c., of the first powers, d_2 a sum of terms which contains squares of differences $dp'_{i|}$, &c. We then have

$$\mathsf{E}z' = c + \mathsf{E}d_2 + \mathsf{E}[d_3 + d_4] + \ldots, \quad dz' = z' - \mathsf{E}z' = D + \frac{1}{N}k,$$

if we put

$$D = d_1 + d_2 + \ldots \text{ and } \frac{1}{N}k = -\{\mathsf{E}d_2 + \mathsf{E}[d_3 + d_4] + \ldots\}$$

Hence we obtain

$$\mathsf{E}\{dz'\}^{2h} = \mathsf{E}D^{2h} + \binom{2h}{1}\frac{1}{N}k\mathsf{E}D^{2h-1} + \ldots + \frac{1}{N^{2h}}k^{2h}$$

$$\mathsf{E}\{dz'\}^{2h+1} = \mathsf{E}D^{2h+1} + \binom{2h+1}{1}\frac{1}{N}k\mathsf{E}D^{2h} + \ldots + \frac{1}{N^{2h+1}}k^{2h+1}$$

and easily satisfy ourselves that $\mathsf{E}\{dz'\}^{2h}$ does not contain any terms with lower powers of $\frac{1}{N}$ than $\left(\frac{1}{N}\right)^h$ and that $\mathsf{E}\{dz'\}^{2h+1}$ does not contain any terms with lower powers of $\frac{1}{N}$ than $\left(\frac{1}{N}\right)^{h+1}$. Hence we have only to stick to the rule, if stopping the extensions in series of powers of $dm'_{1|0}$, $d\mu''_{1|0}$, &c., at terms of even order, in order to obtain correctly the expansion of terms of $\mathsf{E}u'$ in powers of $\frac{1}{N}$ up to the term of order $\left(\frac{1}{N}\right)$-degree equal to the half of the ordinal number of the last retained term in $dm'_{1|0}$, $d\mu''_{1|0}$, &c. But this rule must be strictly observed. The occa-

sional neglect of it has loaded some investigations by English statisticians with errors by which the results aimed at have been seriously vitiated.

F. The expansion in series of powers of $\frac{1}{N}$ presupposes that terms of higher order in $\frac{1}{N}$ are sufficiently small for the total of the terms following the last one retained to be neglected. Strictly speaking, this can be strongly substantiated only by convergency tests or by the analysis of the retained terms. However, up to the present statisticians have not gone so far. Mostly we confine ourselves to the computation and the consideration of the term of order $\frac{1}{N}$. The term of order $\left(\frac{1}{N}\right)^2$ is calculated in but few cases. Here lies an unrestrictedly wide scope for activity for the mathematical statistician, in which beginners can combine the training of their own powers with the achievement of scientifically valuable results in a highly profitable manner. Yet every improvement, however far-reaching in the expansion in series of powers $\frac{1}{N}$, will be halted by the demand that the number of trials must be large. This method can by no means be employed in cases where N is not large. Sometimes the bounds of its applicability become still narrower. Often the series in question proceed not in powers of $\frac{1}{N}$ but in powers of $\frac{1}{Np_{i|}}$ or of $\frac{1}{Np_{i|j}}$, &c., so that it must be demanded that the number of repetitions of various possible values of X and Y and their different combinations is large, and not only the total number of trials. With greater probabilities $p_{i|}$, $p_{i|j}$, &c., however, the difference hardly comes into consideration. But with smaller values of probabilities it may become considerable. And within the realm of the so-called law of small numbers, i.e. with probabilities so small that the product Np with infinitely increasing N tends to a finite limit, the method

of expansion in series of powers of $\frac{1}{N_p}$ completely fails. If, in such cases, we wish to form judgement upon the reliability of our method of estimation of *a priori* magnitudes on the basis of empirical material, we must look for other methods of inquiry.

In order to overcome difficulties which involve the computation of mathematical expectations of quotients, we can under some circumstances employ the following method with advantage. Let Z' and W' be two integral rational functions of empirical values of variables X and Y, and let the function U' employed for the estimation of the *a priori* magnitude interesting to us be equal to the quotient of Z' by $W' : U' = \frac{z'}{w'}$. If C denotes a constant, then we have identically

$$\frac{1}{w'} = \frac{1}{c} - \frac{w' - c}{cw'} = \frac{1}{c} - \frac{w' - c}{c^2} + \frac{(w' - c)^2}{c^2 w'} = \ldots$$

and $$\mathsf{E}\frac{z'}{w'} = \frac{\mathsf{E}z'}{c} - \frac{\mathsf{E}z'[w' - c]}{c^2} + \frac{1}{c^2}\mathsf{E}\frac{z'[w' - c]^2}{w'}.$$

If all possible values of the quotient $\frac{z'}{w'}$ are positive or equal 0 (or ≤ 0), which is true in many cases under consideration, then, assuming that all possible values ≥ 0,

$$\mathsf{E}\frac{z'}{w'}[w' - c]^2 > 0$$

and consequently

$$\text{(I)} \qquad \mathsf{E}\frac{z'}{w'} > \frac{1}{c}\mathsf{E}z' - \frac{1}{c^2}\mathsf{E}z'[w' - c].$$

On the other hand, if the quotient $\frac{z'}{w'}$ cannot be larger than a definite constant, k^2—for instance, not larger than 1—we obtain

$$\mathsf{E}\frac{z'}{w'} < \frac{1}{e}\mathsf{E}z' - \frac{1}{c^2}\mathsf{E}z'[w' - c] + \frac{k^2}{c}\mathsf{E}[w' - c]^2$$

or with $k^2 = 1$

$$\text{(II)} \qquad \mathsf{E}\frac{z'}{w'} < \frac{1}{c}\mathsf{E}z' - \frac{1}{c^2}\mathsf{E}z'[w' - c] + \frac{1}{c}\mathsf{E}[w' - c]^2.$$

Hence, we obtain an upper and lower limit, between which the sought-for value of $\mathsf{E}\frac{z'}{w'}$ must lie with any number of trials.

The indefinite constant C contained in the inequalities can be selected in different ways. If we put $c = Ew'$, then the inequalities are reduced to

$$\mathsf{E}\frac{z'}{w'} > \frac{Ez'}{Ew'} - \frac{1}{[Ew']^2}\mathsf{E}z'[w' - \mathsf{E}w']$$

$$\mathsf{E}\frac{z'}{w'} < \frac{Ez'}{Ew'} - \frac{1}{[Ew']^2}\mathsf{E}z'[w' - \mathsf{E}w'] + \frac{1}{[Ew']^2}\mathsf{E}[w' - \mathsf{E}w']^2.$$

The value of $\mathsf{E}\frac{z'}{w'}$ will thus be confined within limits the difference of which equals $\frac{1}{[Ew']^2}\mathsf{E}[w' - \mathsf{E}w']^2$ and consequently is of order $\frac{1}{N}$. If a higher precision is aimed at, then the substitution $\frac{1}{w'} = \frac{1}{Ew'} - \frac{w' - Ew'}{w'Ew'}$ can be repeated, thus obtaining the generalized inequalities

$$\text{(I')} \quad \mathsf{E}\frac{z'}{w'} > \frac{Ez'}{Ew'} + \sum_{h=1}^{2t-1} \frac{(-1)^h}{[Ew']^{h+1}}\mathsf{E}z'[w' - \mathsf{E}w']^h$$

$$\text{(II')} \quad \mathsf{E}\frac{z'}{w'} < \frac{Ez'}{Ew'} + \sum_{h=1}^{2t-1} \frac{(-1)^h}{[Ew']^{h+1}}\mathsf{E}z'[w' - \mathsf{E}w']^h + \frac{1}{[Ew']^{2t}}\mathsf{E}[w' - \mathsf{E}w']^{2t}.$$

The difference between the upper and lower limits now comes to $\frac{1}{[Ew']^{2t}}\mathsf{E}[w' - \mathsf{E}w']^{2t}$ and is thus of order $\left(\frac{1}{N}\right)^t$.

The value of C can also be selected for each of the inequalities, separately, so obtaining in the first inequality a possibly small upper limit for $\mathsf{E}\frac{z'}{w'}$ and in the second a possibly high lower limit for $\mathsf{E}\frac{z'}{w'}$. The inequality (I) is

thus reduced to $$\mathsf{E}\frac{z'}{w'} > \frac{[Ez']^2}{Ez'w'},$$

and the inequality (II) takes the form

$$\mathsf{E}\frac{z'}{w'} < 1 - \frac{[Ew' - Ez']^2}{E(w')^2 - Ez'w'}.$$

The inequalities (I′) and (II′) allow us to ascertain the terms up to $\left(\frac{1}{N}\right)^{t-1}$ inclusive, in the expansion of $\mathsf{E}\frac{z'}{w'}$ in increasing powers of $\frac{1}{N}$, and at the same time to gain an idea of the magnitude of the error, committed if the terms of higher orders in $\frac{1}{N}$ are neglected. However, this fundamental privilege must be paid for by such laborious computation that it demands still greater patience in the computer than other methods of inquiry. If our aim is to expand $\mathsf{E}\frac{z'}{w'}$ in increasing powers of $\frac{1}{N}$, it is more practical to proceed from the expansion of $\frac{z'}{w'}$ in increasing powers of differences $dp'_{i|j}$, $dm'_{1|0}$, &c.

5. Finally, we should like to consider the parameter already mentioned repeatedly, $\frac{\delta}{\sqrt{p_{1|}p_{2|}p_{|1}p_{|2}}}$, which, when both variables can take only two different values each, may be considered as a correlation coefficient as well as a correlation ratio and also as a mean square contingency. Its empirical counterpart created in the usual way, $\frac{\delta'}{\sqrt{p'_{1|}p'_{2|}p'_{|1}p'_{|2}}}$, can likewise be represented as $r'_{1|1}$ and as $\eta'_{y|x}$ and also as φ'. Hence, there are ways open for determination of $\mathsf{E}\frac{\delta'}{\sqrt{p'_{1|}p'_{2|}p'_{|1}p'_{|2}}}$ and of the standard error of $\frac{\delta'}{\sqrt{p'_{1|}p'_{2|}p'_{|1}p'_{|2}}}$ apart from direct treatment of this function of empirical values; this applies to the formulae relating to $\mathsf{E}r'_{1|1}$, $\sigma_{r'_{1|1}}$, $\mathsf{E}\eta'_{y|x}$, $\sigma_{\eta'_{y|x}}$, $\mathsf{E}\varphi'$ and $\sigma_{\varphi'}$. It is only necessary to insert in

these formulae the values of the parameters m, μ, and r, which we obtain under the assumption that X and Y are able to take two different values each, viz. to confine ourselves to those parameters to which we shall return later,

$$\mu_{2|0} = p_{1|}p_{2|}[x_1 - x_2]^2 \qquad \mu_{0|2} = p_{|1}p_{|2}[y_1 - y_2]^2$$

$$r_{1|1} = \frac{\delta}{\sqrt{p_{1|}p_{2|}p_{|1}p_{|2}}}$$

$$r_{4|0} = \frac{1}{p_{1|}p_{2|}} - 3 \qquad r_{0|4} = \frac{1}{p_{|1}p_{|2}} - 3$$

$$r_{3|1} = r_{4|0}r_{1|1} \qquad r_{1|3} = r_{0|4}r_{1|1}$$

$$r_{2|2} = \frac{(p_{1|} - p_{2|})(p_{|1} - p_{|2})\delta}{p_{1|}p_{2|}p_{|1}p_{|2}} + 1 = r_{1|1}\frac{(p_{1|} - p_{2|})(p_{|1} - p_{|2})}{\sqrt{p_{1|}p_{2|}p_{|1}p_{|2}}} + 1.$$

If these values are put in the formulae for $\mathsf{E}r'_{1|1}$ and $\sigma_{r'_{1|1}}$ we obtain:

$$\mathsf{E}\frac{\delta'}{\sqrt{p'_{1|}p'_{2|}p'_{|1}p'_{|2}}} = \frac{\delta}{\sqrt{p_{1|}p_{2|}p_{|1}p_{|2}}}$$

$$\left\{1 + \frac{1}{N}\left[1 + \frac{(p_{1|} - p_{2|})(p_{|1} - p_{|2})\delta - \frac{1}{2}(p_{1|}p_{2|} + p_{|1}p_{|2})}{4p_{1|}p_{2|}p_{|1}p_{|2}}\right] + \ldots\right\}$$

$$\sigma^2 = \frac{1}{N}\left\{\left[1 + \frac{1}{2}r^2_{1|1}\right]\left[1 + \frac{(p_{1|} - p_{2|})(p_{|1} - p_{|2})\delta}{p_{1|}p_{2|}p_{|1}p_{|2}}\right] - \right.$$

$$\left. - \frac{3}{4}r^2_{1|1}\left[\frac{p_{1|}p_{2|} + p_{|1}p_{|2}}{p_{1|}p_{2|}p_{|1}p_{|2}} - 6\right]\right\} + \ldots$$

With $\delta = 0$, $r_{1|1} = 0$, and $\mathsf{E}r'_{1|1} = 0$: hence, the systematic error of estimation disappears in this case. The standard error of $r'_{1|1}$ thus comes within the values of our approximation to $\sqrt{\frac{1}{N}}$ and, as we know, equals precisely

$$\sqrt{\frac{1}{N-1}}.$$

Further, the systematic error disappears—at least within the limits of our approximation—if the probabilities $p_{1|}$, $p_{2|}$, $p_{|1}$ and $p_{|2}$ all equal $\frac{1}{2}$. The standard error of $r'_{1|1}$ then comes, in the first approximation, to

$$\sqrt{\frac{1 - r^2_{1|1}}{N}} = \sqrt{\frac{1 - 16\delta^2}{N}}.$$

When $p_{1|} = p_{2|} = \frac{1}{2}$, but the difference $p_{|1} - p_{|2}$ is different from 0, the systematic error is negative. But it may also be positive, for instance, if δ is positive and both the differences $p_{1|} - p_{2|}$ and $p_{|1} - p_{|2}$ are positive and sufficiently large.

The estimate of the *a priori* value of $\frac{\delta}{\sqrt{p_{1|}p_{2|}p_{|1}p_{|2}}}$, according to the empirical value of $\frac{\delta'}{\sqrt{p'_{1|}p'_{2|}p'_{|1}p'_{|2}}}$, can thus be connected with a systematic underestimation as well as with a systematic overestimation, but under some circumstances it may be exact.

§ 5

The closer consideration of mathematical expectations of functions of empirical values, taken as points of departure for the estimation of *a priori* indices, has shown that the estimation is, almost without exception, affected by a systematic error which inclines sometimes to one side, sometimes to the other, so that the value sought is over- or underestimated. In valuing these theoretical results from the practical standpoint of the search for statistical correlation, the presence of a systematic error of estimation is of less consequence than its possible magnitude, particularly in comparison with the standard error of the function of empirical values concerned. Only when the systematic error of estimation is of the order of magnitude of a standard error does its neglect seriously influence the results of the inquiry.

Considered from this standpoint, the results of theoretical examinations of systematic errors turn out to be relatively favourable. By increasing the number of trials the systematic errors of estimation rapidly decrease. They might be of great consequence only in a small number of trials. But with small numbers of trials standard errors are also great. A fact of particular importance is that the systematic error of estimation is as a rule of order $\frac{1}{N}$ and the

standard error of order $\sqrt{\frac{1}{N}}$, so that with a fairly considerable number of trials N, the systematic error may be considered small in comparison with the standard error. Let us, for instance, examine more closely the mathematical expectation of the empirical coefficient of correlation in the case of normal correlation. The systematic error in this case is, within our approximation, always negative and equal to $\frac{r_{1|1}[1 - r^2_{1|1}]}{2N}$, and the standard deviation equals, to the first approximation, $\frac{1 - r^2_{1|1}}{\sqrt{N}}$. With a not very large number of trials the systematic error of estimation can be considerable: for instance, N must be greater than 20 in order that we may be certain that the second decimal place is not affected by the systematic error of estimation. But the standard deviation is always considerably greater: the ratio of the systematic error to the standard deviation equals $\frac{r_{1|1}}{2\sqrt{N}}$. When $N = 20$ the *a priori* correlation coefficient must exceed $\frac{3}{4}$ if the standard deviation is not to have a significant *first* decimal place.

Hence, the theoretical analysis of the systematic error leads to the reassuring conclusion that the usage observed by statisticians, of the neglect of the systematic error of estimation, appears at most times quite admissible. There are, however, exceptions. It might occur that the systematic error of estimation is of the same order of magnitude as the standard error. In the case of mean square contingency the mathematical expectation of the empirical expression $[\varphi']^2$, comes, for instance, as we have seen, with mutual independence of the variables X and Y to

$$\mathsf{E}[\varphi']^2 = \frac{[k - 1][l - 1]}{N} + \ldots,$$

and in the general formula for $\sigma^2_{[\varphi']^2}$, with mutual independence of X and Y, the term of order $\frac{1}{N}$ disappears, so that

the standard error—$\sigma_{[\varphi']^2}$—is in this case not of order $\sqrt{\frac{1}{N}}$, but of the same order in $\frac{1}{N}$ as the systematic error of estimation. We are led to a similar conclusion by the consideration of $\mathsf{E}[\eta'_{y|x}]^2$ and $\sigma[\eta'_{y|x}]^2$, when Y is uncorrelated with X and $m^{(i)}_{|1} = m_{0|1}$, $\eta_{y|x} = r_{1|1} = 0$. We obtain in fact

$$\mathsf{E}[\eta'_{y|x}]^2 = \frac{1}{N}\left[\frac{1}{\mu_{0|2}}\sum_i - \mu^{(i)}_{|2}\,1\right] + \ldots$$

and in the formula for $\sigma^2_{[\eta'_{y|x}]^2}$ the term of order $\frac{1}{N}$ disappears, so that the standard error—$\sigma[\eta'_{y|x}]^2$—is of the same order of magnitude as $\mathsf{E}[\eta'_{y|x}]^2 - \eta^2_{y|x}$. Further, with mutual independence of the variables X and Y

$$\mathsf{E}[r'_{1|1}]^2 = \frac{1}{N-1} \text{ and } r^2_{1|1} = 0,$$

as we have seen: consequently the systematic error of the estimate of $r^2_{1|1}$ from the empirical value of $[r^1_{1|1}]^2$ equals $\frac{1}{N-1}$. As for the standard error of $[r^1_{1|1}]^2$, the term of order $\frac{1}{N}$ disappears in the general formula for $\sigma^2_{[r_{1|1}]^2}$ if it is assumed that X and Y are mutually independent. Hence, in the case, the systematic error of estimation is likewise of the same order of magnitude as the standard error. Consequently, in the interpretation of small values of $[\varphi']^2$, $[\eta'_{y|x}]^2$ and $[r'_{1|1}]^2$ caution is needed for two reasons: the inquirer must always reckon not only with the magnitude of the standard error but also with the magnitude of the systematic error of estimation. Similarly, the systematic error of estimation must not be ignored if linearity of the regression of Y on X is to be implied (cf. *supra*, § 4, 3, D) from the insignificance of the difference $[\zeta'_{y|x}]^2$.

§ 6

1. The intensity of stochastic connexion between Y and X appears most comprehensively as the extent to which the range of chance fluctuations of Y is reduced by the determination of the value X. We used to judge the

intensity of the connexion between Y and X by the ratio of the mean conditional variance of Y—$\Sigma p_i \mu_{|2}^{(i)}$—to the total variance of Y—$\mu_{0|2}$—, whereby the *a priori* values of conditional variances $\mu_{|2}^{(i)}$ are estimated from chance values of empirical variances $\mu_{|2}^{(i)'}$. We can, however, gain a certain idea of the intensity of the connexion without referring to the values of magnitudes $\mu_{|2}^{(i)'}$ and $\mu_{|2}^{(i)}$.

Assume that the number of trials does not exceed the number of possible values of X and that accidentally X takes different values in all trials. Then only one value of Y corresponds to every value of X observed. Hence, no single one of the conditional variances of Y can be even roughly estimated. Yet, under some circumstances, the empirical material thus shaped permits of our judging the intensity of the connexion with certainty.

Of course, under such conditions neither the presence of a functional relationship between Y and X nor the appearance of a more or less loose stochastic connexion appear impossible. However whimsically scattered single points may lie, laws of functional relationship which give the observed ordinate values corresponding to the observed abscissal values can nevertheless always be put forward. On the other hand, the supposition always appears obvious that the empirically chance values of Y deviate more or less from the conditional mathematical expectations in question and the true line of regression of Y on X does not coincide exactly with that indicated by the values observed. Both possibilities must not be decided against without further consideration. But the probabilities of different forms of the law of dependence do not remain unaffected by the appearance of the line which represents the observed values : a given succession of individual points permits the appearance of the possible shape of the true regression and the intensity of the connexion as mutually conditional to a certain degree. The assumption of non-correlation of Y with X can, for instance, almost exclude the assumption of a greater intensity of individual points

clinging clearly enough to a line which is not parallel to the X-axis. Similarly, the assumption of linear or parabolic regression with certain forms of empirical material can accompany a high degree of connexion. Under some circumstances values of X and Y co-ordinated with each other can limit the selection of a partially plausible form of the law of dependence so that we arrive at a nearly cogent conclusion: when, for instance, with a sufficiently larger number of trials, all values of Y are strictly proportional to the corresponding values of X, we will infer the presence of a linear functional relationship with a confidence born of practical certainty even in the case where to each value of X there corresponds only one single value of Y.

The inquirer may infer considerably more precise conclusions in relation to the intensity of connexion if he is able to proceed from definite assumptions with regard to the shape of the true line of regression, just as he will come to more precise conclusions about the form of the true regression if he may take definite assumptions as a basis with regard to the intensity of connexion. In the reduction of measurements affected by accidental errors the problem is often put as follows: the precision of the measurements is held as known to some extent and it is demanded of the regression equation that it should represent the observations with probability sufficiently great under the supposition of the assumed precision.

2. The form of the line of regression of Y on X and the intensity of connexion between Y and X do not mutually condition each other (cf. Chap. IV, § 6). The totality of the *a priori* values of $m_{|1}^{(i)}$ does not throw any light on the intensity of connexion; but it presents instead an exhaustive picture of the regression of Y on X. On the other hand, the totality of the empirical $m_{|1}^{(i)}$-values does not represent any reliable picture of regression; but it allows us to create a certain idea of the regression as well as of the intensity of connexion. The problem is similar to that which we considered just before (cf. *supra*, § 6, 1): the more that individual values of Y, and also their arithmetic

mean which appear as $m_{|1}^{(i)'}$-values, cling more closely to the true line of regression, the more intensive is the connexion between Y and X; if the series of $m_{|1}^{(i)'}$-values is given, then not all assumptions in relation to regression and to the intensity of connexion appear equally consistent with each other. The inquirer is thus enabled to arrive at comparatively reliable and precise judgements with regard to regression and intensity of connexion, rejecting assumptions with apparently small probability. If the inquirer is thus able to proceed from definite assumptions with regard to regression, then the reliability of his estimation of the intensity increases. If he possesses certain knowledge about the intensity of connexion, then his judgement with regard to regression can be made with greater certainty.

3. We stand on firmer ground in estimating the intensity of stochastic connexion between the variables if we are able to rely on the consideration not only of $m_{|1}^{(i)'}$-values but also of $m_{1|}^{(j)'}$-values at the same time. As we have seen (Chap. IV, § 6), acquaintance with both the *a priori* regression equations allows us to estimate more or less precisely the intensity of the connexion and the cognizance of empirical $m_{|1}^{(i)}$- and $m_{1|}^{(j)'}$-values in its turn permits us to ascertain with more or less security equations of *a priori* lines of regression. By the computation of empirical regression coefficients $b'_{1|}$ and $b'_{|1}$ in the equations of lines, which give the best possible fit to the empirical lines of regression (cf. Chap. V, § 3), we can, as $b'_{1|}b'_{|1} = [r'_{1|1}]^2$, assess the value of the empirical coefficient of correlation and from the latter—$r'_{1|1}$—we may infer in the usual way (cf. *supra*, § 4, 3, A) the value of the *a priori* correlation coefficient $r_{1|1}$ which may hold under known limitations (cf. Chap. IV, § 4, 3) as a measure of intensity of connexion.

Hence, if we have at hand the totality of $m_{|1}^{(i)'}$-values and of $m_{1|}^{(j)'}$-values we can gain a conception of the stochastic connexion between variables sufficient for many purposes. Neither the standard deviation nor the systematic error of estimation of the *a priori* magnitudes in question can, how-

ever, be ascertained on the basis of the knowledge of these values alone: even the true form of lines of regression can be ascertained with greater reliability and exactness only if the totality of co-ordinated values of X and Y is at our disposal. Thus it is preferable, by far, to have the empirical material in the form of a detailed correlation table and not comprehensively a series of $m_{11}^{(i)\prime}$-values and of $m_{11}^{(j)\prime}$-values: only the well-planned refinement of original observations allows us to squeeze from the results of measurements all they are able to disclose about the stochastical connexion between variables. In the publication of empirical material this should always be borne in mind.

§ 7

As our subject-matter is the presentation of the theory of the methods applied to the investigation of stochastic connexion between two chance variables, I should like to point out very briefly that in the search for stochastic connexion between more than two chance variables analogous methods of inquiry are mainly employed. Some new notions are, however, added. Apart from conditional laws of distribution we have to deal also with conditional laws of dependence; we have to investigate conditional coefficients of correlation, conditional correlation ratios, &c. Besides lines of regression within the field of contemplation of an inquirer there appear also regression surfaces as well as formations of three, and more, dimensions. Of particular importance in correlation inquiry and for the comprehensive presentation of associations between three, or more, stochastically connected chance variables, are the highly valuable notions of multiple and partial correlation ratios and correlation coefficients. The new problems, to which the consideration of more than two stochastically connected chance variables leads, offer a highly theoretical and practical interest. Their treatment would require so much room, however, that it would appear more appropriate to set it aside for the time being.

CHAPTER VII

STOCHASTIC SUPPOSITION OF THE MEASUREMENTS OF CORRELATION

§ 1

THE methods of estimation of *a priori* magnitudes on the basis of empirical values of variables analysed in Chapter VI proceed from the assumption that the joint frequency-distribution does not change from trial to trial and that individual trials are mutually independent (cf. Chap. VI, § 3, 1). When the joint frequency-distribution changes and individual trials are not independent, we must not rely on the formulae considered in Chapter VI without further consideration.

Assume that the frequency-distribution remains constant but the individual trials are connected with each other in a manner corresponding to the scheme of drawing from a closed urn without replacing the tickets drawn. Let the total number of tickets in the turn be A ; each ticket is marked with two numbers, one in black ink specifying the value of X, and one in red ink the corresponding value of Y. If one draws N tickets from the urn in such a way that the extracted tickets are always replaced in the urn before the next extraction takes place, then the *a priori* coefficients which characterize the joint frequency-distribution of black and red numbers on the tickets in the urn must be estimated by a known method on the basis of numbers, marked on the extracted tickets , for instance, for the estimation of the *a priori* coefficient of correlation the formulae of § 4, 3, A, of Chapter VI must be employed :

$$(\mathrm{I})\begin{cases} \mathsf{E}r'_{1|1} = r'_{1|1} + \frac{1}{N}\left\{\frac{1}{4}r_{2|2}r_{1|1} + \frac{3}{8}r_{1|1}[r_{4|0} + r_{0|4}] - \right. \\ \qquad\qquad \left. - \frac{1}{2}[r_{3|1} + r_{1|3}]\right\} + \ldots \\ \sigma^2_{r'_{1|1}} = \frac{1}{N}\left\{r_{2|2}\left[1 + \frac{1}{2}r^2_{1|1}\right] - r_{1|1}[r_{3|1} + r_{1|3}] + \right. \\ \qquad\qquad \left. + \frac{1}{4}r^2_{1|1}[r_{4|0} + r_{0|4}]\right\} + \ldots \end{cases}$$

But if the extracted tickets are not replaced the mathematical expectation and the variance of the empirical correlation coefficient cannot be computed by means of these formulae. The formulae that hold good in this case must be derived anew.

The derivation can be carried out in the same manner as in the case of mutual independence of trials. One proceeds, as above (cf. Chap. VI, § 4), from the expansion of $r'_{1|1}$ in increasing powers of differences, $dp'_{i|j}$, $dp'_{i|}$, &c., or of the differences $dp'_{i|j}$, $dp'_{i|}$, &c., or of the differences $dm'_{1|1}$, $d\mu''_{1|1}$, &c., then one turns in a similar way to the mathematical expectations. The mathematical expectations of different powers of $dp'_{i|j}$, $dp'_{i|}$, $dm'_{1|1}$, $d\mu''_{1|1}$, &c., however, have different forms in the two cases. If in the expansion of series of $\mathsf{E}r'_{1|1}$ those values of the mathematical expectations are inserted which correspond to the scheme of unreplaced tickets, one obtains instead of the formulae (I):

$$(\mathrm{II})\begin{cases} \mathsf{E}r'_{1|1} = r_{1|1} + \frac{A-N}{A-1}\frac{1}{N}\left\{\frac{1}{4}r_{2|2}r_{1|1} + \frac{3}{8}r_{1|1}[r_{4|0} + r_{0|4}] - \right. \\ \qquad\qquad \left. - \frac{1}{2}[r_{3|1} + r_{1|3}]\right\} + \ldots \\ \sigma^2_{r'_{1|1}} = \frac{A-N}{A-1}\frac{1}{N}\left\{r_{2|2}\left[1 + \frac{1}{2}r^2_{1|1}\right] - r_{1|1}[r_{3|1} + r_{1|3}] + \right. \\ \qquad\qquad \left. + \frac{1}{4}r^2_{1|1}[r_{4|0} + r_{0|4}]\right\} + \ldots \end{cases}$$

We see that the systematic error of estimation as well as the variance of $r'_{1|1}$ is reduced through the presence of such a connexion between the trials, in the ratio of $A - N$

to $A - 1$ in comparison with the case of non-connected trials : the mathematical expectation of $r'_{1|1}$ differs less from the *a priori* value of $r'_{1|1}$ and the range of chance fluctuations is likewise less. Thanks to the connexion of trials, one can be more confident in finding with the aid of the empirical-chance value $r'_{1|1}$ a good approximation to the true magnitude of $r'_{1|1}$.

Assume, on the other hand, that the extracted ticket is replaced in the urn before the next extraction takes place and that simultaneously another ticket is put in the urn bearing the same numbers entered in black and in red ink, respectively, as in the ticket just extracted. I shall call this scheme ' extractions with additions '. If individual trials are connected in such a manner, neither the formulae (I) nor the formulae (II) can be employed in estimating the *a priori* correlation coefficient. The expansion of series of $r'_{1|1}$ in increasing powers of differences $dp'_{i|j}$, $dp'_{i|}$, &c., or of differences $dm'_{1|1}$, $d\mu''_{1|1}$, &c., still holds, but, again, the mathematical expectations of different powers of differences take another form. If the values of the mathematical expectations concerned are inserted in the general expansion of the series we obtain :

$$(\text{III}) \begin{cases} \mathsf{E}r'_{1|1} = r_{1|1} + \dfrac{A+N}{A+1}\dfrac{1}{N}\left\{\dfrac{1}{4}r_{2|2}r_{1|1} + \dfrac{3}{8}r_{1|1}[r_{4|0} + r_{0|4}] - \right. \\ \qquad\qquad \left. - \dfrac{1}{2}[r_{3|1} + r_{1|3}]\right\} + \ldots \\ \sigma^2_{r'_{1|1}} = \dfrac{A+N}{A+1}\dfrac{1}{N}\left\{r_{2|2}\left[1 + \dfrac{1}{2}r^2_{1|1}\right] - r_{1|1}[r_{3|1} + r_{1|3}] + \right. \\ \qquad\qquad \left. + \dfrac{1}{4}r^2_{1|1}[r_{4|0} + r_{0|4}]\right\} + \ldots \end{cases}$$

In contrast to the scheme of unextracted tickets the systematic error of estimation as well as the variance are in this case increased, viz. to a first approximation, both the ratio of $A + N$ to $A + 1$. Hence, if the trials are connected in this manner, then the mathematical chance-value of $r'_{1|1}$ aims with less accuracy a point which lies at a greater distance from the goal.

Other assumptions in relation to the connexion of trials as well as dispensing with the assumption that the joint frequency-distribution remains constant, would lead to new formulae. Hence, the estimation has to allow in every case for the appropriate magnitude of the standard error and for systematic error.

The same holds good for correlation ratios, for the mean square contingency, &c. At every inference from empirical values to *a priori* magnitudes we must always take into account the stochastic suppositions involved in forming estimates from empirical values.

§ 2

Since in inferring *a priori* magnitudes from empirical values under difficult conditions different methods must be used, the question must be asked, how can the statistician decide which formulae or which stochastic suppositions he should take as a basis for his computations in each individual case? The inquirer has to deal with the same question in an investigation of an individual chance variable. If we wish to infer from the variance of black numbers on the tickets extracted from an urn the variance of black tickets on the tickets remaining in the urn, we are dependent in extractions with replacement on the formula

$$\mathsf{E}\mu'_{0|2} = \frac{N-1}{N}\mu_{0|2},$$

in extractions without replacement on the formula

$$\mathsf{E}\mu'_{0|2} = \frac{A}{A-1}\frac{N-1}{N}\mu_{0|2},$$

and in extractions with addition on the formula

$$\mathsf{E}\mu'_{0|2} = \frac{A}{A+1}\frac{N-1}{N}\mu_{0|2}.$$

Accordingly, should the *a priori* variance—$\mu_{0|2}$—be estimated on the basis of observation, then it is not sufficient to compute the empirical variance—$\mu_{0|2}$. It is also necessary to have some knowledge of the realization of empirical values.

In the examination of chance variables the statistician uses the computation of the divergence coefficient to support the judgement of stochastic suppositions. As is known, this method is due to W. Lexis. Originally the limits of its application were drawn rather narrowly by Lexis, who had exclusively in mind statistical numbers referring to frequencies or to known functions of frequencies. L. von Bortkiewicz then transferred the Lexis method to frequency-distributions of any form. Fundamentally the Lexis method consists in the introduction of a special criterion under the name of a ' *Divergence Coefficient* ' which assumes the value 1, if the frequency-distribution of variables remains constant at all trials, and the trials are mutually independent. Statistical series which satisfy this condition Lexis calls ' normally stable '. If the value of the divergence coefficient calculated on the empirical material lying before us diverges more from unity than is consistent with a range of chance fluctuations in question, then the series cannot be normally stable : either the frequency-distribution does not remain constant or the trials are not independent or both. We may, on the other hand, assume, with certain reservations, that the stochastic suppositions of normal stability are present, if the computations result in a value of the divergence coefficient which approximately equals 1.

We may make use of the same method at the investigation of three or more stochastically connected chance variables. Let us denote the *stochastic connexion* as *normally stable* if the joint frequency-distribution remains constant at all trials and they are mutually independent. A criterion can be constructed for this case which corresponds to the Lexis-Bortkiewicz divergence coefficient, which can also be called a divergence coefficient. If all suppositions of normal stability are fulfilled, this divergence coefficient comes to 1. If it diverges significantly from 1, then the suppositions of the normal stability are not fulfilled : either joint frequency-distribution changes from trial to trial or the trials are connected. If the divergence coefficient approaches closely

enough to 1, one may, with the same reservation as in the case of variable, assume that the stochastic suppositions of normal stability are verified, the joint frequency-distribution remains constant and the trials are mutually independent.

§ 3

From the assumptions that the joint frequency-distribution remains constant at all trials and that the individual trials are mutually independent follows immediately:

$$\mathsf{E}[x^{[f]\prime} - m_{1|0}]^c[y^{[f]\prime} - m_{0|1}]^d = \mu_{c|d}$$

$$\mathsf{E}[x^{[f]\prime} - m_{1|0}]^c[y^{[g]\prime} - m_{0|1}]^d = \{\mathsf{E}[x^{[f]\prime} - m_{1|0}]^c\}\{\mathsf{E}[y^{[g]} - m_{0|1}]^d\} = \mu_{c|0}\mu_{0|d}(f \neq g)$$

and then

$$\mathsf{E}\left\{\frac{1}{N}\sum_{f=1}^{N}[x^{[f]\prime} - m_{1|0}]\,[y^{[f]\prime} - m_{0|1}]\right\} - \mu_{1|1}$$

$$\mathsf{E}\left\{\frac{1}{N-1}\sum_{f=1}^{N}[x^{[f]\prime} - x'_0]\,[y^{[f]\prime} - y'_0]\right\} = \mu_{1|1}.$$

Let us split up the N trials into r-series of n-trials each and denote by $x_0^{[h]\prime}$ and $y_0^{[h]\prime}$ the arithmetic means of the chance values of X and of Y respectively for the hth series. If all N-trials are considered as a whole we find:

$$\mathsf{E}\left\{\frac{1}{rn-1}\sum_{f=1}^{rn}[x^{[f]\prime} - x'_0]\,[y^{[f]\prime} - y'_0]\right\} = \mu_{1|1}.$$

On the other hand, if we proceed from the consideration of r-series, we obtain

$$\mathsf{E}\left\{\frac{1}{r-1}\sum_{h=1}^{r}[x_0^{[h]\prime} - x'_0]\,[y_0^{[h]\prime} - y'_0]\right\} = \frac{1}{n}\mu_{1|1}.$$

If we define the divergence coefficient Q as

$$Q = \frac{\dfrac{n}{r-1}\displaystyle\sum_{h=1}^{r}[x_0^{[h]\prime} - x'_0]\,[y_0^{[h]\prime} - y'_0]}{\dfrac{1}{rn-1}\displaystyle\sum_{f=1}^{rn}[x^{[f]\prime} - x'_0]\,[y^{[f]\prime} - y'_0]},$$

then it is as easy to prove as in the case of a chance variable that $EQ = 1$ if all suppositions of a normal stability hold good, i.e. if the joint frequency-distribution remains constant at all trials and the trials are mutually independent. The analogy with the Lexis-Bortkiewicz divergence coefficient comes clearly to light in so far as the above expression for Q is converted into the Lexis-Bortkiewicz divergence coefficient, if it is assumed that the value of Y always coincides with the corresponding value of X.

§ 4

If the correlation is normally stable, then computations which have in view estimations of *a priori* magnitudes on the basis of empirical value of variables have an easily explicable sense: one gains a rough idea of certain coefficients which comprehensively characterize the constant remaining joint frequency-distribution. If, however, the stability is not normal, then not only the manner of inferring the *a priori* magnitudes from empirical values changes, but also the meaning of the results obtained may be altered.

In the consideration of non-normal stability we have to distinguish two cases: the stability can be abnormal, because the trials are connected, although the joint frequency-distribution remains constant; it can be abnormal because the joint frequency-distribution changes from trial to trial. It is true that in the first case the estimation of the *a priori* magnitudes on the basis of empirical material is difficult because the trials are not independent; but if we surmount the difficulties and come to sufficiently based presumptive values of *a priori* magnitudes, then their meaning is the same as in the case of normally stable correlation: we gain a rough idea of coefficients which comprehensively characterize the constant remaining joint frequency-distribution. For instance, reviewing the kinds of connected trials considered above—the scheme of extractions without replacement and the scheme of extraction with additions: the computation of empirical correlation

coefficient leads, in both cases, to the value of the *a priori* correlation coefficient which has exactly the same meaning as the measure of the intensity of connexion between black and red numbers on the tickets in the urn as in the case of replacement: only the standard deviations and the systematic errors are different in the three cases. If, however, the stability is abnormal, because the joint frequency-distribution changes, then the calculated presumptive values have no longer the same meaning because the coefficients ascertained by its means do not characterize any definite joint frequency-distribution, but on the contrary, refer to the totality of changeable joint frequency-distributions, and must be understood as a mean value of *a priori* coefficients which characterize individual joint frequency-distribution. The computations remain the same in all the cases: the empirical correlation coefficient, for instance, is always calculated by the same formula. But what is found by its calculation varies in meaning: at times the meaning of the number calculated can be precisely seized, at other times its subject is more or less vague.

The dependence of the sense of the results aimed at by the statistician refinement of the empirical material on the kind of stochastic suppositions made is more important than the differences in the magnitudes of the standard error and of the systematic error of estimation. The statistician must always endeavour to look thoroughly into the stochastic suppositions of the empirical materials which lie before him, looking as carefully as possible into the facts themselves, into their causal conditions as well as into the technique of the collection of data, and he must also, as far as it is possible, support his examination by the computation of divergence coefficients. It is no less important in the investigation of stochastic connexion than when it is a matter of a single chance variable.

CHAPTER VIII

OBJECT AND VALUE OF CORRELATION MEASUREMENT

§ 1

1. What does the inquirer gain from the computation of correlation coefficients, correlation ratios, &c. ? What are the advantages of these ' mathematical ' methods of inquiry over non-mathematical ones ?

First of all : in the more precise framing of judgement. Even an investigator who does not undertake any measurements cannot abstain from judging with the intensity and other measurable properties of the relations between the statistical series which he compares. His judgements, however, remain highly subjective.

Without much calculation one notices, for instance, that some of the series which one has to deal with are closely similar, whereas others are rather different. If one has the instinct for such eye-estimates such judgement of the intensity of the relation can even turn out fairly precisely. Some members of my Seminar practised making such estimates for a while, as a sort of statistical game. Those among them who were particularly efficient at it went so far as to be able to read off correctly the first decimal place of the correlation coefficient from the graphical representation of two series. A mechanized method, however, has the same advantages for such estimates here as elsewhere. Likewise one need not at first compute the arithmetic mean precisely in order to form a rough idea of the level of fluctuation of the numbers of a statistical series : an attentive consideration of individual values allows the well-trained eye not only to grasp without any long calculation whether

the average of a series is higher or lower than that of another one, but even perhaps to state the approximate magnitude of the difference. In a similar way one is able to form a judgement without having calculated the standard deviation whether one series shows a smaller or greater fluctuation than another. Such judgements, however, are rather unsafe as a rule. An eye-estimate can be deceptive. There are not many people who can unconcernedly rely upon their capacity for measurement. Two investigators with the same series of numbers before them will often come to contradictory conclusions, and in such cases as nobody likes to admit that he has less ability than his neighbour for visual estimation each of them will think he is in the right and reject the other's judgement as subjective. The verification by means of ' mathematical ' methods is then the only means of deciding the controversy. The well-known American statistician, W. C. Mitchell, has published a monograph on business cycles which is the most thorough analysis of the statistical material in this field.* Mitchell has renounced the application of mathematical methods in this work and has sought to base his conclusions with regard to the fluctuations of the series of numbers and the intensity of the relations between them on graphical considerations. Another American statistician, B. W. King, has taken the trouble to verify Mitchell's judgement by precise calculations.† Most of them proved to be correct, but in some cases material corrections had to be made. Estimates of fluctuation in this case turned out far more exact than judgements on the intensity of association.

Furthermore, one may compare the rudimentary ' non-mathematical ' methods of computing lines of regression with the self-contained systems of ideas of the mathematical theory of correlation, which culminates in the simultaneous

* W. C. Mitchell, *Business Cycles* (Memoirs of the University of California, Vol. 3) ; 1913.

† B. W. King, ' A Study of Mitchell's Inquiries into Prices ' (*Quarterly Journal of Economics*, Vol. XXXI, 1917).

consideration of the most plausible regression equation and the correlation ratio. The non-mathematician also forms a notion of whether values of one variable increase or decrease on the average with the growth of the values of another one, and is even able to gain an idea of the rate of increase or decrease when the form of the regression curve does not deviate too significantly from linearity. However, he works with rather vague notions and with still vaguer ideas of the suppositions on which the method he is employing depends; his quantitative judgements suffer by uncertainty and inevitable subjectivity and he is not in a position to attach due consideration to the disturbing influence of chance fluctuations; either he is too confident or, disillusioned, he begins to be too cautious in his conclusions. The mathematical statistician, on the contrary, is in a position to make a more precise estimate of the reliability of his conclusions by the computation of the relevant standard error. The regression equation allows him to compute beforehand the expected values of Y which correspond to different values of X, and the correlation ratio of Y on X gives him the average measure of the range of chance fluctuations, of which Y is still capable after the determination of the value of X. When he has to deal with several stochastically connected variables he can ascertain by the computation of correlation ratio of T on X, on Y, on Z, &c., the relative importance of individual factors for the prediction of the value of T, and again he can at the same time determine by this computation how far the fluctuations of T may be attributed to the influence of other factors. If the correlation ratio of T on X, Y, Z, is equal to 1, this means that T is functionally related to X, Y, Z, so that the value of T is determined with certainty by the given values of X, Y, Z. And in the cases where the correlation ratio of T on X, Y, Z, though not exactly equal to 1, is yet not greatly divergent from 1; the finished investigation may be considered a success because the clue has been found to those factors which substantially deter-

mine the value of T. In this case the regression equation of T on X, Y, Z, permits one to predict the value of T, not indeed with certainty as in the case of functional relationship but still with the less uncertainty, the more the correlation ratio of T on the factors used for the computation approach the maximum value of 1. Particularly for scientific prognosis in the field of non-functional relationships—for example, for the prediction of trade movements—the methods built on the computation of regression equations and the corresponding correlation ratios are an absolutely decisive step on the way to a rational solution of the problem.

2. In fixing notions with precision and differentiating with exactitude the different coefficients in question, a further advantage of the mathematical methods of inquiry becomes apparent. The non-mathematician who avoids measures can, at best, get a vague idea that the association of the series in question is in one case somewhat more intense than in another one. What he has in mind as a measure of intensity when forming his judgement cannot be exactly ascertained. The mathematical statistician, on the contrary, is able to distinguish different coefficients, to interpret their numerical values in a manner corresponding to their meaning, and to draw essential conclusions from the comparison of the numerical values of different coefficients. If, for instance, numerical values of the correlation ratio of Y on X and of the correlation coefficient between X and Y differ more than is compatible with the magnitude of the standard error of the difference, then it is a sure sign that the regression of Y on X cannot be linear. When the numerical values of coefficients allow of a reasonable interpretation the determination of the numerical values is in itself a scientifically valuable result. Measures of the intensity of relationship are usually so constructed that they attain a definite value—mostly to the value 0—in the case of mutual independence and another value—mostly the value 1—in the case of a connexion of maximum intensity. It is sought to graduate the scale of values which lie between

0 and 1 as far as possible, so that they may represent as closely as possible the degree of intensity of the relation. Of course, the conceptions of what is really being graduated are often not too clear to the discoverers of new coefficients, and the meaning of the 'independence' corresponding to the value 0 as well as the connexion of maximum intensity corresponding to the value 1 are sometimes likewise not so easy to define with precision. But even if it is not possible to prove that the graduations of the scale represent correspondingly measured grades of the character considered—as, for instance, is the case in the example of throwing white and red dice which we considered, where the numerical value of the correlation coefficient was equal to the ratio of the remaining dice to the total number of dice (cf. Chap. IV, § 4, 3)—yet as a rule the numerical value of the coefficient bears the meaning that greater values of the coefficient point to greater intensities of the character to be measured. That measurements thus made are not valueless is confirmed by the scientific practice in all branches of science. What doubles if the temperature increases from 5° C. to 10° C. ? Or does something rather change in the ratio of 41 to 50 which is obtained if the temperatures are measured in degrees Fahrenheit ?

3. The determination of the numerical values of coefficients has an importance apart from the obvious interpretation of them. For the calculation of the standard errors enables us to delimit the range of chance fluctuations to be taken into consideration. Even in deciding, in the first place, whether there is any association at all between the phenomena under examination, this is of great importance. Through the chance fluctuations of empirical values, even when there is no association at all, the existence of a more or less clear connexion can be simulated. The mathematical statistician compares the deviation of the empirical numerical value of the chosen coefficient from its independence value with the appropriate standard error. In this way he arrives at an objectively based judgement, as to whether the devia-

tion can, under all circumstances, be held to be sufficiently large to lie no longer within the range of chance fluctuations ; in fact, whether the parallelism between the numerical series under observation is not, after all, traceable to the influence of chance causes. This way is barred to the non-mathematical statistician who avoids measures of association. His task can be solved only by roundabout procedures which, in spite of considerable expenditure of industry and ingenuity, usually do not lead to quite satisfactory results. There are similar difficulties to be surmounted when judging whether the relation in one case is more intense than in another, &c. : here mathematical methods of inquiry always show to advantage against non-mathematical ones, which avoid the use of measures of association.

§ 2

1. The advantage of the mathematical method becomes most apparent if the investigation is not confined to a simple pair of stochastically connected variables but has to deal with a greater number of measurements. The statistician who avoids measures of association, when considering several series simultaneously, seldom gets beyond the conclusion that some of the series before him show a direct and some an inverse relation. Finer distinctions in the form of the relation escape his eye. If, however, methodical measurements are made, further analysis of the results is possible. We compute the numerical values of coefficients chosen for series of statistical observation which are set under different circumstances ; from the differences and from the agreements between the values obtained, inferences can be made which provide a deeper insight into the relationships observed and which help to throw light upon their meaning. We may suppose, for instance, that the computed coefficients of correlation exhibit a clearly apparent regularity of distribution in space of sequence in time, that larger and smaller values of coefficients of correlation are grouped on the map in a way which allows the

emergence of clearly outlined zones of more or less intense connexion between the phenomena under examination, or that in the chronological order of coefficients of correlation an intensity of connexion which varies systematically—i.e. a continued increase or decrease—comes to light. Properly speaking, every such objective regularity—in space and in time—in the form of computed sequences of correlation coefficients, correlation ratios, &c., is in the first instance but a new enigma, but through the consideration and deciphering of a system of such enigmas the inquirer gains the key to the solution of the most important among the tasks before him, the interpretation of the true quality of the relationships manifested in his numerical series (cf. Chap. II, § 2). Particularly among the meteorologists who like drawing on maps lines of equal intensity of association similar to isobars and isotherms to supplement the analysis of their observational material the consideration of regularities in the spatial distribution of correlation coefficients is firmly established. But also in other branches of science similar methods are beginning to be used to a steadily increasing extent with equally good success. Let us consider more closely by an example in what way systematic measurements of intensity of connexion can contribute to the augmentation of attainable results.

2. No other country possesses such reliable and ample statistics of the consumption of spirits as Russia during the time of the State monopoly of the sale of spirits. I have attempted in my Seminar to turn this splendid material to the best scientific account. The central point of our investigation was the question of the influence of the harvest upon the consumption of spirits in Russia. There are many publications with widely contradictory opinions on that question in Russia. The inquiry carried out in my Seminar by Miss M. Winogradowa,* who knew how to

* M. M. Winogradowa *The Consumption of Spirits and the Harvest in Russia* (Investigations of Students of the Economic Faculty of the Polytechnic Institute of Petrograd, No. 17); 1916 (Russian).

make exemplary use of the statistical material of the Board of Monopoly by means of extensive computation of correlation coefficients, clarified the problem. From this inquiry, so instructive in methodology and content, I will borrow some illustrations of the value of skilful inferences from serial correlations in answering the questions under examination.

The first question which had to be elucidated was, whether the harvest had influenced at all markedly the consumption of spirits in Russia—as was assumed by most of the investigators who had worked on this problem, but which was categorically denied by one of the greatest authorities on theoretical economics in Russia, W. Dmittrieff in his *Critical Investigation of the Consumption of Alcohol in Russia.* Dmittrieff believes, rather, that he can show that the fluctuations of the consumptions of spirits in Russia are traceable not to fluctuations of the harvest but to those of the industrial trade cycle: the harvest as the dominant factor in Russian economics is already played out, he thinks, owing to industrial development. Dmittrieff seeks to support his thesis by analysis of the statistical material and to corroborate it by a psychological theory of alcoholism. The statistics he uses—the yield of the harvest and the consumption of spirits in European Russia in the individual years of the period considered by him—disclose no really marked influence of the level of harvest upon the consumption of spirits. Now it was obvious that this might be owing to the imperfection of the method of investigation he had chosen. The consumption during one calendar year comes under the influence of two harvests: the consumption in the last month of the year under the influence of the same year's harvest and the consumption in the first month of the year under the influence of the previous year's harvest. Hence, if successive harvests turn out differently, then their effects are levelled on the average of the calendar year. Again, considerations which refer to the whole territory of European Russia likewise appear for reasons of the

same sort little suited to bring out clearly the influence of harvest upon the consumption: the fluctuations of the whole harvest are chiefly determined by the yield in the southern parts of Russia, whilst the fluctuations of the total consumptions are to a great extent determined by the consumption in the northern provinces also, and the harvest of the north can show a very different result from that in the south. Statistical sources from the time of monopoly allow a more refined method of inquiry: there are monthly returns of accounts of the consumption of spirits in individual provinces. Thus one can create consumption-years by additions of the twelve-monthly consumptions which are influenced by the same harvest, and compare the fluctuations of the consumption with that of the harvest in the individual provinces. As the State monopoly of sale was not carried out simultaneously in all parts of Russia, Miss Winogradowa confined the calculation to the nineteen provinces for which there was material for longer periods. The picture of the geographical distribution of the correlation coefficients shows characteristic features. We see on the map a zone of high correlation coefficients of over 0·75: these are the most important corn-producing provinces, with the province of Samara at the top with the correlation coefficient of + 0·98; to this group also belong Ufa and Orenburg, provinces adjacent to Samara, as well as the strip of southern provinces extending westwards: Jekaterinoslaw, Cherson, Poltawa, Kiew, Kamenez-Podolsk. Close to this zone, towards the north, there are transition zones with lower but still not insignificant positive correlation coefficients. On the other hand, for the provinces farther north one arrives at considerably lower correlation coefficients (with the exception of the province of Olonez), and in some points even at negative ones.

The geographical regularity in the distribution of values of correlation coefficients which measure the intensity of the connexion between the harvest and the consumption of spirits gives promise of important information with regard

to the kind of relationship. This insight can be deepened if we compute correlation coefficients for separate months and for individual provinces. Taken as a whole, these figures manifest further striking regularities. The provinces of the first zone, which are characterized by the very high positive correlation coefficients of the year's consumption with the harvest, are distinguished by the fact that the high positive correlation of the consumption of spirits with the harvest persists throughout all the months of the harvest-year: for instance, in the province of Samara no correlation coefficient falls below + 0·5 for the period commencing October until September of the following year. In those provinces where the year's consumption is not so closely connected with the harvest, a more detailed inspection shows, however, that there are parts of the year which show a greater connexion than others. Even in those provinces where the year's consumption is negatively correlated, often a positive correlation can be ascertained for a greater or smaller portion of the harvest-year. Accordingly the result of the harvest has a different effect in different provinces: at times its influence is more durable, at times it drops more quickly; sometimes it makes its appearance early, sometimes not before late autumn. It must be noted that the connexion is nowhere very intense during the months of harvesting, but sometimes after this period it begins to occur more intensely everywhere. In the province of Samara, for instance, the correlation coefficient does not fall below 0·80 for the months October to April, whereas it remains under 0·5 during the months July, August, September. This regularity is of decisive importance in the judgement of Dmittrieff's psychological constructions. It follows, then, that the higher consumption of spirits in the years of good harvest can neither be explained by the physiological needs of the peasants and agricultural workers, more than usually exhausted from the gathering of a rich harvest, nor explained by the ' unanimated ' mood of the country people, aroused by the favourable results of

the harvest—as was suggested by Dmittrieff. The influence of the harvest on the consumption of spirits must clearly be rooted in something quite different.

If we now ask what is the real basis of the connexion between the consumption of spirits and the results of the harvest, a connexion which varies so much in various regions and at different seasons, the most obvious answer is: it is based on the fact that the population disposes of the more abundant means accruing from a good harvest by spending it on spirits. This answer was generally held to be satisfactory until Dmittrieff made an attempt to refute it. That this explanation is the right one appears highly plausible in view of the results of the correlations we have considered: differences in the intensity of connexion which we have noticed can very well be explained by the fact that the result of a harvest does not everywhere play the same part in the peasant's household and that the harvest does not occur everywhere at the same time. In such regions as Samara, the results of a harvest almost completely determine the year's income, and moreover, thanks to its southern position, the harvesting is carried out and the crop brought to market earlier. In northern Russia the harvest, even in good years, does not suffice to cover all household needs. The household depends to a greater or smaller extent on the income from other sources (industrial activity, labour in forests, &c.). Here the influence of a good harvest cannot last long, and, moreover, it may be complicated by variations in the income from other sources.

This explanation, in itself quite plausible, of the influence of the harvest on the consumption of spirits is irrefutably confirmed by the further results of Miss Winogradowa's investigations. If this hypothesis is correct the decisive influence on the consumption of spirits must be not the quantity but the money-value of the harvest. It will be more strongly marked, therefore, during those months in which that part of the harvest not reserved for the house-

hold itself is taken to market. In order to trace these factors Miss Winogradowa has computed correlation coefficients for a series of provinces, between the consumption of spirits during individual months and the money-receipts for those agricultural products which, in the regions in question, play a particularly large part. It was everywhere proved that the connexion between the consumption of spirits and yield is most intense for the yield of those agricultural products which in the months and region in question are brought to market by the peasants. It has been shown, for instance, for the province of Smolensk, where flax and hemp are cultivated chiefly for sale and the main kind of grain—rye—is grown more for household needs, that in August the consumption of spirits is, in fact, positively correlated with the money-value of the rye yield (the correlation coefficient $= + 0{\cdot}3$), but that by September the rye-yield ceases to influence the consumption ; on the other hand, in September a higher correlation of the consumption of spirits with the value of flax and hemp is observable, to which oats and potatoes are added in October and November, and in December the value of the yield of flax and hemp fibre.

The authoress has had the insight to complete this extremely clear picture of the relationship between yield and consumption of spirits by further refinements. For instance, the statistical sources allow of differentiation of the sale of spirits in receptacles of different sizes : in large gallon bottles, in wine-bottles of usual size, in half-sizes of the latter and in very small bottles. If one computes the correlation coefficients between the sale of spirits in different-sized bottles and the harvest results, a clear increase of coefficients for the larger bottles will appear : in the province of Ufa the correlation coefficient comes, for example, to $+ 0{\cdot}8$ for the gallon bottles, to $+ 0{\cdot}7$ for the ordinary wine-bottle size, to $+ 0{\cdot}6$ for half-sizes of the latter and to $+ 0{\cdot}4$ for the very small bottles. This regularly increasing intensity of the connexion as we proceed from smaller

to larger receptacles gives a deep insight into the relationships. Consumption of spirits in Russia can be placed in two classes: people to whom alcohol stimulation has become a necessity seek to satisfy their wants as far as their means permit and they indulge in the desired drop as regularly as possible; these regular and habitual drinkers procure their brandy in smaller bottles. But very much brandy, particularly in the country, is consumed on occasions of parish festivities, fairs, and particularly weddings. For this purpose the peasant procures brandy in larger receptacles. Thus from the above values of the correlation coefficients it can be concluded that the relationship between harvest-yield and consumption of spirits is essentially traceable to the fact that in the years of good yields the peasants more richly endowed with money, celebrate their festivities and weddings with particular plenty and, with regard to the weddings, in greater number. The relationship between harvest-yields and frequency of weddings on the one hand, and between the frequency of weddings and the consumption of spirits on the other, can be also proved directly. The curves show a remarkable parallel, particularly when the sale in gallon bottles is taken into account.

In this way an important connecting-link is inserted between the fluctuations due to harvest-yields of the purchasing power of the peasants and the consumption of spirits. Methods which avoid measurement of the intensity of the connexion were in most cases considered by us merely able to ascertain the presence of an undifferentiated relationship: there were no remarkable differences between the sale in large and in small bottles, between the consumption of spirits in different months and in different provinces, between the quantity and the money-value of the yield, &c. Only through mathematical methods can colour and life be brought into the monotonous grey picture of non-mathematicians. Only the methodical measurements of the intensity allows us to gain a deeper insight into the nature of the relationship.

Of course, it does not follow that it is sufficient to calculate according to well-known formulae correlation ratios, correlation coefficients, &c., in series in order to obtain a deeper insight into the relationships in question. The modern theory of correlation puts at the inquirer's disposal a rich assortment of refined tools. He who understands how to handle them skilfully is able to extract from his figures much which would otherwise remain concealed. But the value of the production will never be determined by the property of tools alone. A few hasty strokes sketched out on a scrap of paper by the hand of a master calls into being a picture which surpasses in workmanship many a multicoloured painting executed with the greatest diligence. In order that the methods worked out from the theory of correlation may lead to a deeper understanding of the relationships under examination, the statistician who employs them must be master of his problem. It is not enough to be familiar with the technical tools ; he must be familiar with the subject of the investigation as well, and he must have complete command of his material. He must possess the ability to adapt the technique of his investigation to the end pursued and to the possibilities before him. A routine-like mechanical reliance on ready-made prescriptions leads, even when the most complicated formulae are employed and the most precise calculations are carried out, to an unproductive waste of time and energy and to the accumulation of numerical values which are but little likely to enrich our essential knowledge.

APPENDIX

CHAPTER I

§ 2, C. Denoting by d_i the difference between the ith values of the series X and Y and bearing in mind that

$$\sum_{x=1}^{n} x = \frac{n(n+1)}{2}, \quad \sum_{x=1}^{n} x^2 = \frac{n(n+1)(2n+1)}{6},$$

in the case when the series Y is arranged in decreasing order of magnitude, we have

$$\sum_{i=1}^{n} d_i^2 = [1-n]^2 + [2-(n-1)]^2 + \ldots + [h-(n-h+1)]^2 + \ldots + [(h-1)-2]^2 + [n-1]^2 =$$

$$= \sum_{h=1}^{n} [h-(n-h+1)]^2 = \sum_{h=1}^{n} [2h-(n+1)]^2 =$$

$$= 4\sum_{h=1}^{n} h^2 - 4(n+1)\sum_{h=1}^{n} h + n(n+1)^2 = \frac{n(n^2-1)}{3}.$$

When the order of the series Y is independent of that of the series X, every member of the series Y may occur with every member of the series X with the same probability $\frac{1}{n}$. For the square of the difference d_i we obtain, in this case, the expected value

$$d_i^2 = \frac{1}{n}[i-1]^2 + \frac{1}{n}[i-2]^2 + \ldots + \frac{1}{n}[i-(i-1)]^2 +$$

$$+ \frac{1}{n}[i-i]^2 + \frac{1}{n}[i-(i+1)]^2 + \ldots + \frac{1}{n}[i-(n-1)]^2 +$$

$$+ \frac{1}{n}[i-n]^2 = \frac{1}{n}\sum_{h=1}^{i-1} h^2 + \frac{1}{n}\sum_{h=1}^{n-i} h^2 =$$

$$= \frac{1}{n}\left\{\frac{(i-1)i(2i-1)}{6} + \frac{(n-i)(n-i+1)(2n-2i+1)}{6}\right\} =$$

$$= \frac{1}{6}[2n^2+3n+1] - (n+1)i + i^2.$$

Whence through the summation from $i = 1$ to $i = n$ it follows that

$$\sum_{i=1}^{n} d_i^2 = \frac{n(2n^2 + 3n + 1)}{6} - \frac{n(n+1)^2}{2} + \frac{n(n+1)(2n+1)}{6} = \\ = \frac{n(n^2-1)}{6}.$$

CHAPTER IV

§ 1, 2. Apart from the formulae mentioned in the text the following identities should be noted, to which continual reference will be made:

$$p_{i|} = \sum_{j=1}^{l} p_{ij} \qquad p_{|j} = \sum_{i=1}^{k} p_{ij}$$

$$1 = \sum_{i} p_{i|} = \sum_{j} p_{|j} = \sum_{i}\sum_{j} p_{ij}$$

$$1 = \sum_{i} p_{i|}^{(j)} = \sum_{j} p_{|j}^{(i)}.$$

§ 2, 1. Noting that when $k = l = 2p_{1|} = p_{1|1} + p_{1|2}$, $p_{1|} + p_{2|} = 1$, &c., it is easy to see that

$$\delta = p_{1|1} - p_{1|}p_{|1} = p_{1|} - p_{1|2} - p_{1|}[1 - p_{|2}] = -[p_{1|2} - p_{1|}p_{|2}] = p_{2|2} - p_{2|}p_{|2} = -[p_{2|1} - p_{2|}p_{|1}].$$

As $p_{1|1} + p_{1|2} + p_{2|1} + p_{2|2} = 1$, we have

$$\delta = p_{1|1} - p_{1|}p_{|1} = p_{1|1}[p_{1|1} + p_{1|2} + p_{2|1} + p_{2|2}] - \\ - [p_{1|1} + p_{1|2}]\,[p_{1|1} + p_{2|1}] = p_{1|1}p_{2|2} - p_{1|2}p_{2|1}.$$

For the value of the mean square contingency, when both the variables can assume only two different values each, we easily obtain, bearing in mind the above identities,

$$\varphi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{[p_{ij} - p_{i|}p_{|j}]^2}{p_{i|}p_{|j}} = \delta^2\left\{\frac{1}{p_{1|}p_{|1}} + \frac{1}{p_{1|}p_{|2}} + \frac{1}{p_{2|}p_{|1}} + \frac{1}{p_{2|}p_{|2}}\right\} = \\ = \frac{\delta^2}{p_{1|}p_{2|}p_{|1}p_{|2}}.$$

Appendix

§ 3, 1. Noting that $m_{|1}^{(i)} = \sum_j p_{|j}^{(i)} y_j$, we have

$$\sum_i p_{i|} m_{|1}^{(i)} = \sum_i \sum_j p_{i|} p_{|j}^{(i)} y_j = \sum_i \sum_j p_{i|j} y_j = \sum_j p_{|j} y_j = m_{0|1}.$$

Taking into consideration

$$\sum_i \sum_j p_{i|j} [y_j - m_{0|1}] [m_{|1}^{(i)} - m_{0|1}] =$$

$$= \sum_i \{p_{i|} [m_{|1}^{(i)} - m_{0|1}] \sum_j p_{|j}^{(i)} [y_j - m_{0|1}]\} = \sum_i p_{i|} [m_{|1}^{(i)} - m_{0|1}]^2,$$

we have

$$\sum_i p_{i|} \mu_{|2}^{(i)} = \sum_i \sum_j p_{i|} p_{|j}^{(i)} [y_j - m_{|1}^{(i)}]^2 =$$

$$= \sum_i \sum_j p_{i|j} [(y_j - m_{0|1}) - (m_{|1}^{(i)} - m_{0|1})]^2 =$$

$$= \sum_i \sum_j p_{i|j} [y_j - m_{0|1}]^2 - \sum_i p_{i|} [m_{|1}^{(i)} - m_{0|1}]^2 =$$

$$= \mu_{0|2} - \sum_i p_{i|} [m_{|1}^{(i)} - m_{0|1}]^2.$$

§ 3, 1. As $m_{f|g} = \sum_i \sum_j p_{i|j} x_i^f y_j^g$ and $m_{f|0} m_{0|g} = \sum_i \sum_j p_{i|} p_{|j} x_i^f y_j^g$, we have $m_{f|g} - m_{f|0} m_{0|g} = \sum_i \sum_j [p_{i|j} - p_{i|} p_{|j}] x_i^f y_j^g$. From $m_{f|g} - m_{f|0} m_{0|g} = 0$ we obtain $\sum_i \sum_j [p_{i|j} - p_{i|} p_{|j}] x_i^f y_j^g = 0$. This relation can hold good at all positive integral values of f and g only if all differences $p_{i|j} - p_{i|} p_{|j}$ are equal to zero, i.e. if the variables are mutually independent.

§ 3, 2, A.

$$\sum_i p_{i|} x_i^h m_{|1}^{(i)} = \sum_i p_{i|} x_i^h \sum_j p_{|j}^{(i)} y_j = \sum_i \sum_j p_{i|j} x_i^h y_j = m_{h|1}.$$

§ 3, 2, B. Noting that $\sum_j p_{|j}^{(i)} = 1$ and that obviously

$$m_{|1}^{(i)} - m_{0|1} = \sum_j p_{|j}^{(i)} y_j - \sum_j p_{|j}^{(i)} m_{0|1} = \sum_j p_{|j}^{(i)} [y_j - m_{0|1}],$$

we have

$$\sum_i p_{i|} [x_i - m_{1|0}]^h [m_{|1}^{(i)} - m_{0|1}] = \sum_i \sum_j p_{i|j} [x_i - m_{1|0}]^h$$

$$[y_j - m_{0|1}] = \mu_{h|1}.$$

Appendix

§ 3, 2, C. If we write the regression equation in normal co-ordinates in the form

$$\mathfrak{M}_{|1}^{(i)} = c_{|0} + c_{|1}\mathfrak{X}_i + c_{|2}\mathfrak{X}_i^2 + \ldots + c_{|f}\mathfrak{X}_i^f$$

and note that

$$\sum_i p_{i|}\mathfrak{X}_i^h = \frac{1}{\sigma_x^h}\sum_i p_{i|}[x_i - m_{1|0}]^h = \frac{\mu_{h|0}}{\sigma_x^h} = r_{h|0}$$

$$\sum_i p_{i|}\mathfrak{M}_{|1}^{(i)}\mathfrak{X}_i^h = \frac{1}{\sigma_x^h\sigma_y}\sum_i p_{i|}[x_i - m_{1|0}]^h\,[m_{|1}^{(i)} - m_{0|1}] =$$

$$= \frac{\mu_{h|1}}{\sigma_x^h\sigma_y} = r_{h|1}$$

and that $r_{0|0} = 1$, $r_{1|0} = r_{0|1} = 0$, $r_{2|0} = r_{0|2} = 1$, we obtain for the determination of the coefficients c the linear equations

$$0 = c_{|0} + c_{|2} + \ldots + r_{f|0}c_{|f}$$

$$r_{1|1} = c_{|1} + r_{3|0}c_{|2} + \ldots + r_{f+1|0}c_{|f}$$

$$r_{2|1} = c_{|0} + r_{3|0}c_{|1} + r_{4|0}c_{|2} + \ldots + r_{f+2|0}c_{|f},\ \text{&c.}$$

For the case of a parabola of the second degree we have

$$0 = c_{|0} + c_{|2}$$

$$r_{1|1} = c_{|1} + r_{3|0}c_{|2}$$

$$r_{2|1} = c_{|0} + r_{3|0}c_{|1} + r_{4|0}c_{|2}.$$

§ 3, 4. Since

$$\frac{\partial\sum_i p_{i|}[m_{|1}^{(i)} - A_{|0} - A_{|1}x_i]^2}{\partial A_{|0}} = -2\sum_i p_{i|}[m_{|1}^{(i)} - A_{|0} - A_{|1}x_i] =$$

$$= -2[m_{0|1} - A_{|0} - A_{|1}m_{1|0}]$$

and

$$\frac{\partial\sum_i p_{i|}[m_{|1}^{(i)} - A_{|0} - A_{|1}x_i]^2}{\partial A_{|1}} = -2\sum_i p_{i|}x_i[m_{|1}^{(i)} - A_{|0} - A_{|1}x_i] =$$

$$= -2[m_{1|1} - A_{|0}m_{1|0} - A_{|1}m_{2|0}],$$

the coefficients $A_{|0}$ and $A_{|1}$ may be determined from the equations

$$A_{|0} + m_{1|0}A_{|1} = m_{0|1}$$

$$m_{1|0}A_{|0} + m_{2|0}A_{|1} = m_{1|1}.$$

Appendix

§ 4, 2. When the regression line is a parabola of the second degree the regression equation can be written in the form

$$\mathfrak{M}_{|1}^{(i)} = r_{1|1}\mathfrak{X}_i + c_{|2}[\mathfrak{X}_i^2 - r_{3|0}\mathfrak{X}_i - 1].$$

It follows that

$$[\mathfrak{M}_{|1}^{(i)}]^2 = r_{1|1}^2\mathfrak{X}_i^2 + 2r_{1|1}c_{|2}\mathfrak{X}_i[\mathfrak{X}_i^2 - r_{3|0}\mathfrak{X}_i - 1] + + c_{|2}^2[\mathfrak{X}_i^2 - r_{3|0}\mathfrak{X}_i - 1]^2.$$

Noting that

$$\sum_i p_{i|}\mathfrak{X}_i[\mathfrak{X}_i^2 - r_{3|0}\mathfrak{X}_i - 1] = r_{3|0} - r_{3|0} = 0$$

$$\sum_i p_{i|}[\mathfrak{X}_i^2 - r_{3|0}\mathfrak{X}_i - 1]^2 = \sum_i p_{i|}\{\mathfrak{X}_i^4 - 2r_{3|0}\mathfrak{X}_i^3 + + [r_{3|0}^2 - 2]\mathfrak{X}_i^2 + 2r_{3|0}\mathfrak{X}_i + 1\} = r_{4|0} - r_{3|0}^2 - 1$$

and that

$$c_{|2} = \frac{r_{2|1} - r_{3|0}r_{1|1}}{r_{4|0} - r_{3|0}^2 - 1},$$

it is easily seen that

$$\eta_{y|x}^2 = \sum_i p_{i|}[\mathfrak{M}_{|1}^{(i)}]^2 = r_{1|1}^2 + c_{|2}^2[r_{4|0} - r_{3|0}^2 - 1] = r_{1|1}^2 + + \frac{[r_{2|1} - r_{3|0}r_{1|1}]^2}{r_{4|0} - r_{3|0}^2 - 1}.$$

§ 4, 3:

$$\sum_i p_{i|}[m_{|1}^{(i)} - M_{|1}^{(i)}]^2 = \sum_i p_{i|}\left[(m_{|1}^{(i)} - m_{0|1}) - \frac{\mu_{1|1}}{\mu_{2|0}}(x_i - m_{1|0})\right]^2 =$$

$$= \sum_i p_{i|}[m_{|1}^{(i)} - m_{0|1}]^2 - 2\frac{\mu_{1|1}}{\mu_{2|0}}\sum_i p_{i|}[x_i - m_{1|0}][m_{|1}^{(i)} - m_{0|1}] +$$

$$+ \frac{\mu_{1|1}^2}{\mu_{2|0}^2}\sum_i p_{i|}[x_i - m_{1|0}]^2 = \mu_{0|2}\eta_{y|x}^2 - \mu_{0|2}r_{1|1}^2.$$

§ 4, 3. Let

$$x = W_1 + W_2 + \ldots + W_m + U_1 + U_2 + \ldots + U_n$$

$$y = W_1 + W_2 + \ldots + W_m + T_1 + T_2 + \ldots + T_l,$$

where $W_1, W_2, \ldots, W_m$; $U_1, U_2, \ldots, U_n$; $T_1, T_2, \ldots, T_l$ are independent chance variables which follow the same law of distribution. Denoting the mathematical expecta-

tion of the variables by m_1 and the variance by μ_2 and noting that

$$\mathsf{E}[W_i - m_1][W_j - m_1] = 0, \quad \mathsf{E}[U_i - m_1][U_j - m_1] = 0,$$
$$\mathsf{E}[T_i - m_1][T_j - m_1] = 0 \quad (j \neq i),$$
$$\mathsf{E}[W_i - m_1][U_j - m_1] = \mathsf{E}[W_i - m_1][T_j - m_1] =$$
$$\mathsf{E}[U_i - m_1][T_j - m_1] = 0 \quad \left(j \gtreqless i\right),$$

we have

$$x - \mathsf{E}x = \sum_{i=1}^{m}[W_i - m_1] + \sum_{j=1}^{n}[U_j - m_1],$$

$$y - \mathsf{E}y = \sum_{i=1}^{m}[W_i - m_1] + \sum_{j=1}^{l}[T_j - m_1],$$

$$\sigma_x^2 = \mathsf{E}[x - \mathsf{E}x]^2 = \sum_{i\ 1}^{m}\mathsf{E}[W_i - m_1]^2 + \sum_{j=1}^{n}\mathsf{E}[U_j - m_1]^2 =$$
$$= [m + n]\mu_2, \quad \sigma_y^2 = [m + l]\mu_2,$$

$$\mathsf{E}\{[x - \mathsf{E}x][y - \mathsf{E}y]\} = \sum_{i=1}^{m}\mathsf{E}[W_i - m_1]^2 = m\mu_2$$

and hence

$$r_{1|1} = \frac{E\{[x - Ex][y - Ey]\}}{\sigma_x\sigma_y} = \frac{m}{\sqrt{[m + n][m + l]}}.$$

§ 5, 2. The parameters

$$r_{f|h} = \frac{1}{2\pi\sqrt{1 - r_{1|1}^2}}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-\frac{\mathfrak{X}^2 - 2r_{1|1}\mathfrak{X}\mathfrak{Y} + \mathfrak{Y}^2}{2[1 - r_{1|1}^2]}}\mathfrak{X}^f\mathfrak{Y}^h d\mathfrak{X}d\mathfrak{Y}$$

can be expressed in different forms. The formulae mentioned in the text are due to K. Pearson and A. W. Young, ' On the Product-Moments of Various Orders of the Normal Correlation Surface of Two Variables ' (*Biometrika*, Vol. XII). The most direct way of obtaining them is by substituting $\mathfrak{X} = r_{1|1}\mathfrak{Y} + Z$ and then expanding $[r_{1|1}\mathfrak{Y} + Z]^f$ in powers of $\mathfrak{Y}$ and Z, when the double integral can be repre-

sented as a product of two single integrals which can be easily calculated from the formula

$$\int_{-\infty}^{+\infty} t^{2n} e^{-ct^2} dt = \frac{1 \cdot 3 \ldots (2n-1)}{2^n} \sqrt{\frac{\pi}{c^{2n+1}}}.$$

From $r_{2f|1|1} = 1 \cdot 3 \cdot 5 \ldots (2f+1) r_{1|1} = r_{2f|2|0} r_{1|1}$ it follows that the regression of Y on X is linear (cf. Chap. IV, § 3, 2, C).

From $r_{2f|0} = 1 \cdot 3 \cdot 5 \ldots (2f-1)$ it can be inferred that the variable X follows the Gauss-Laplace law of distribution.

If Y is held constant, it is easy to see that the distribution of values of X which correspond to a constant value of Y follows the Gauss-Laplace law of distribution and that the conditional variance is equal to $[1 - r_{1|1}^2]\mu_{2|0}$ for any constant value of Y.

The formula $\varphi^2 = \frac{r_{1|1}^2}{1 - r_{1|1}^2}$ is due to K. Pearson, *On the Theory of Contingency and its Relation to Association and Normal Correlation* (Drapers' Company Research Memoirs, Biometric Series, I ; 1904) ; the derivation of the formulae is to be found on pages 7 and 8.

§ 7. When the variables can take only two different values each, we have

$$\mathsf{E}x = p_{1|}x_1 + p_{2|}x_2 \qquad \mathsf{E}y = p_{1|}y_1 + p_{|2}y_2$$

$$x_1 - \mathsf{E}x = p_{2|}[x_1 - x_2] \qquad x_2 - \mathsf{E}x = -p_{1|}[x_1 - x_2]$$

$$y_1 - \mathsf{E}y = p_{|2}[y_1 - y_2] \qquad y_2 - \mathsf{E}y = -p_{|1}[y_1 - y_2]$$

$$\mu_{2|0} = p_{1|}[x_1 - \mathsf{E}x]^2 + p_{2|}[x_2 - \mathsf{E}x]^2 = p_{1|}p_{2|}[x_1 - x_2]^2$$

$$\mu_{0|2} = p_{|1}[y_1 - \mathsf{E}y]^2 + p_{|2}[y_2 - \mathsf{E}y]^2 = p_{|1}p_{|2}[y_1 - y_2]^2$$

$$\begin{aligned}\mu_{1|1} &= \mathsf{E}xy - [\mathsf{E}x]\,[\mathsf{E}y] = (p_{1|1} - p_{1|}p_{|1})x_1y_1 + \\ &+ (p_{1|2} - p_{1|}p_{|2})x_1y_2 + (p_{2|1} - p_{2|}p_{|1})x_2y_1 + \\ &+ (p_{2|2} - p_{2|}p_{|2})x_2y_2 = \delta[x_1y_1 - x_1y_2 - x_2y_1 + x_2y_2] = \\ &= \delta[x_1 - x_2]\,[y_1 - y_2]\end{aligned}$$

$$m^{(1)}_{|1} = \frac{1}{p_{1|}}[p_{1|1}y_1 + p_{1|2}y_2] \qquad m^{(2)}_{|1} = \frac{1}{p_{2|}}[p_{2|1}y_1 + p_{2|2}y_2]$$

$$m_{0|1} = p_{|1}y_1 + p_{|2}y_2$$

$$m^{(1)}_{|1} - m_{0|1} = \frac{1}{p_{1|}}\{[p_{1|1} - p_{1|}p_{|1}]y_1 + [p_{1|2} - p_{1|}p_{|2}]y_2\} =$$

$$= \frac{1}{p_{1|}}\delta[y_1 - y_2]$$

$$m^{(2)}_{|1} - m_{0|1} = -\frac{1}{p_{2|}}\delta[y_1 - y_2]$$

$$\eta^2_{y|x} = \frac{1}{\mu_{0|2}}\sum_{i=1}^{2} p_{i|}[m^{(i)}_{|1} - m_{0|1}]^2 =$$

$$= \frac{\delta^2[y_1 - y_2]^2\left[\frac{1}{p_{1|}} + \frac{1}{p_{2|}}\right]}{p_{|1}p_{|2}[y_1 - y_2]^2} = \frac{\delta^2}{p_{1|}p_{2|}p_{|1}p_{|2}}.$$

CHAPTER V

The formulae of Chapter V are so closely connected with the corresponding formulae of Chapter IV, that their derivation cannot cause any difficulties: it is only necessary to bear in mind the corresponding treatment in Chapter IV.

CHAPTER VI

§ 2, 1. In drawings with replacement the probability P_x of drawing x white balls in N draws is well known to be given by

$$P_x = \frac{N(N-1)\ldots(N-x+1)}{1\cdot 2\ldots(x-1)x}p^x(1-p)^{N-x}$$

and

$$\sum_{x=0}^{N} P_x = [p + (1-p)]^N = 1.$$

Appendix

The mathematical expectation of the number of white balls in N draws can be most simply calculated as follows:

$$\mathsf{E}n = \sum_{x=0}^{N} \frac{N(N-1)\ldots(N-x+1)}{1\cdot 2\ldots(x-1)x} p^x(1-p)^{N-x}x =$$

$$= \sum_{x=1}^{N} \frac{N(N-1)\ldots(N-x+1)}{1\cdot 2\ldots(x-1)} p^x(1-p)^{N-x} =$$

$$= Np\sum_{x=1}^{N} \frac{(N-1)\ldots(N-x+1)}{1\cdot 2\ldots(x-1)} p^{x-1}(1-p)^{N-x};$$

substituting $(t+1)$ for x:

$$\mathsf{E}n = Np\sum_{t=0}^{N-1} \frac{(N-1)\ldots(N-t)}{1\cdot 2\ldots t} p^t(1-\mathrm{p})^{N-1-t} =$$

$$= Np[p + (1-p)]^{N-1} = Np$$

The probability of drawing x white balls in N draws without replacement if the urn contains $B = Ap$ white and $C = A(1-p)$ black balls is equal to

$$P_x = \frac{N(N-1)\ldots(N-x+1)}{1\cdot 2\ldots x}$$

$$\frac{B(B-1)\ldots(B-x+1)C(C-1)\ldots(C-N+x+1)}{A(A-1)\ldots(A-N+1)},$$

where $\sum_{x=0}^{N} P_x = 1$. In the same way as before, we have

$$\mathsf{E}n = \sum_{x=0}^{N} P_x x =$$

$$= \frac{NB}{A}\sum_{x=1}^{N} \frac{(N-1)\ldots(N-x+1)}{1\cdot 2\ldots(x-1)}$$

$$\frac{(B-1)\ldots(B-x+1)C(C-1)\ldots(C-N+x+1)}{(A-1)\ldots(A-N+1)} =$$

$$= Np\sum_{t=0}^{N-1} \frac{(N-1)\ldots(N-t)}{1\cdot 2\ldots t}$$

$$\frac{(B-1)\ldots(B-t)C(C-1)\ldots(C-N+t+2)}{(A-1)\ldots(A-N+1)} = Np;$$

as the sum by which Np is multiplied is precisely the sum of the probabilities of drawing 0, 1, 2, . . ., $N-1$ white balls in $N-1$ draws without replacement, from an urn containing $B-1$ white and C black balls.

Noting that $x^2 = x(x-1) + x$ for draws with replacement we obtain similarly

$$\mathsf{E}n^2 = \sum_{x=0}^{N} p_x x^2 = N(N-1)p^2 + Np = N^2p^2 + Np(1-p),$$

and hence

$$\sigma^2_{\frac{n}{N}} = \mathsf{E}\left[\frac{n}{N} - \mathsf{E}\frac{n}{N}\right]^2 = \mathsf{E}\frac{n^2}{N^2} - p^2 = \frac{1}{N}p(1-p),$$

and for draws without replacement

$$\mathsf{E}n^2 = \frac{N(N-1)B(B-1)}{A(A-1)} + \frac{NB}{A} = \frac{N^2B(B-1)}{A(A-1)} + \\ + \frac{NB}{A}\left[1 - \frac{B-1}{A-1}\right],$$

and hence

$$\sigma^2_{\frac{n}{N}} = \mathsf{E}\frac{n^2}{N^2} - \frac{B^2}{A^2} = \frac{B(A-B)}{A^2(A-1)}\left[\frac{A}{N} - 1\right] = \frac{A-N}{A-1}\frac{1}{N}p(1-p).$$

§ 3, 2. Denoting by $\mathsf{E}^{(h)}_{n_{2|2}}$ the conditional mathematical expectation of $n_{2|2}$ when $n_{1|1} = h$, and by P_h the probability that $n_{1|1}$ assumes the value h, then

$$P_h = \frac{N(N-1)\ldots(N-h+1)}{1\cdot 2\ldots h} p^h_{1|1}[1-p_{1|1}]^{N-h}$$

and $\mathsf{E}n_{1|1}n_{2|2} = \sum_h P_h h \mathsf{E}^{(h)}_{n_{2|2}}$. Assuming that $n_{1|1} = h$, then $N-h$ trials are left for the remaining three combinations of X-values and of Y-values, and the probability that any one of these $N-h$ trials results in the combination of values X_2Y_2 is equal to

$$\frac{p_{2|2}}{p_{2|2}+p_{1|2}+p_{2|1}} = \frac{p_{2|2}}{1-p_{1|1}}.$$ Hence $\mathsf{E}^{(h)}_{n_{2|2}} = \frac{(N-h)p_{2|2}}{1-p_{1|1}}$ and

$$\mathsf{E}n_{1|1}n_{2|2} = \sum_h P_h h(N-h)\frac{p_{2|2}}{1-p_{1|1}} = \frac{p_{2|2}}{1-p_{1|1}}\left\{N\sum_h P_h h - \right.$$

$$\left. - \sum_h p_h h^2\right\} = \frac{p_{2|2}}{1-p_{1|1}}\left\{N^2p_{1|1} - N^2p^2_{1|1} - Np_{1|1}[1-p_{1|1}]\right\} =$$

$$= N(N-1)p_{1|1}p_{2|2}.$$

Appendix

Introducing, for shortness, the notation $N(N-1)\ldots(N-t+1)=N^t$, we obtain similarly

$$\mathsf{E}n_{1|1}n_{2|2}n_{1|2}n_{2|1}=N^4p_{1|1}p_{2|2}p_{1|2}p_{2|1}$$

$$\mathsf{E}n_{1|1}^2n_{2|2}^2=N^4p_{1|1}^2p_{2|2}^2+N^3p_{1|1}p_{2|2}[p_{1|1}+p_{2|2}]+N^2p_{1|1}p_{2|2}$$

$$\mathsf{E}n_{1|2}^2n_{2|1}^2=N^4p_{1|2}^2p_{2|1}^2+N^3p_{1|2}p_{2|1}[p_{1|2}+p_{2|1}]+N^2p_{1|2}p_{2|1}$$

and hence

$$\begin{aligned}N^4\mathsf{E}[\delta']^2&=\mathsf{E}n_{1|1}^2n_{2|2}^2+\mathsf{E}n_{1|2}^2n_{2|1}^2-2\mathsf{E}n_{1|1}n_{2|2}n_{1|2}n_{2|1}=\\&=N^4[p_{1|1}p_{2|2}-p_{1|2}p_{2|1}]^2+N^3\{p_{1|1}p_{2|2}[p_{1|1}+p_{2|2}]+\\&\quad+p_{1|2}p_{2|1}[p_{1|2}+p_{2|1}]\}+N^2[p_{1|1}p_{2|2}+p_{1|2}p_{2|1}]=\\&=N(N-1)\{N^2\delta^2+N[p_{1|1}p_{2|2}(p_{1|1}+p_{2|2})+p_{1|2}p_{2|1}\\&\quad(p_{1|2}+p_{2|1})-5\delta^2]+[p_{1|1}p_{2|2}+p_{1|2}p_{2|1}-2p_{1|1}p_{2|2}\\&\quad(p_{1|1}+p_{2|2})-2p_{1|2}p_{2|1}(p_{1|2}+p_{2|1})+6\delta^2]\}\end{aligned}$$

$$\begin{aligned}\sigma^2_{\frac{N}{N-1}\delta'}&=\mathsf{E}\Big[\frac{N}{N-1}\delta'\Big]^2-\delta^2=\\&=\frac{1}{N-1}\{[p_{1|1}p_{2|2}(p_{1|1}+p_{2|2})+p_{1|2}p_{2|1}(p_{1|2}+p_{2|1})-\\&\quad-4\delta^2]+\frac{1}{N}[(p_{1|2}+p_{2|1})\delta-(p_{1|1}+p_{2|2})\delta+6\delta^2]\}.\end{aligned}$$

§ 4, 1, B. As the law of distribution of X is the same at all trials and the individual trials are mutually independent,

$$\mathsf{E}\frac{[x^{[f]\prime}-x_0']^2}{\Sigma_1^2}=\mathsf{E}\frac{[x^{[d]\prime}-x_0']^2}{\Sigma_1^2}$$

and

$$\mathsf{E}\frac{[x^{[f]\prime}-x_0'][x^{[d]\prime}-x_0']}{\Sigma_1^2}=\mathsf{E}\frac{[x^{[h]\prime}-x_0'][x^{[c]\prime}-x_0']}{\Sigma_1^2}.$$

Hence we have

$$\mathsf{E}\frac{[x^{[f]\prime}-x_0'][x^{[d]\prime}-x_0']}{\Sigma_1^2}=\frac{1}{N-1}\mathsf{E}\frac{[x^{[f]\prime}-x_0']\sum\limits_{d\neq f}[x^{[d]\prime}-x_0']}{\Sigma_1^2}=$$

$$=\frac{1}{N-1}\mathsf{E}\frac{[x^{[f]\prime}-x_0']\Big\{\sum\limits_{d\neq f}[x^{[d]\prime}-x_0']+[x^{[f]\prime}-x_0']-[x^{[f]\prime}-x_0']\Big\}}{\Sigma_1^2}=$$

$$=-\frac{1}{N-1}\mathsf{E}\frac{[x^{[f]\prime}-x_0']^2}{\Sigma_1^2}.$$

Similarly we find

$$\mathsf{E}\frac{[y^{[f]\prime}-y'_0]\,[y^{[d]\prime}-y'_0]}{\Sigma_2^2}=-\frac{1}{N-1}\mathsf{E}\frac{[y^{[f]\prime}-y'_0]^2}{\Sigma_2^2}.$$

Substituting the above values in

$$\mathsf{E}[r'_{1|1}]^2=\mathsf{E}\left\{\sum_{f=1}^{N}\frac{[x^{[f]\prime}-x'_0]\,[y^{[f]\prime}-y'_0]}{\Sigma_1\Sigma_2}\right\}^2=$$

$$=\mathsf{E}\left\{\sum_{f=1}^{N}\frac{[x^{[f]\prime}-x'_0]^2[y^{[f]\prime}-y'_0]^2}{\Sigma_1^2\Sigma_2^2}\right\}+$$

$$+\mathsf{E}\left\{\sum_{f=1}^{N}\sum_{d\neq f}\frac{[x^{[f]\prime}-x'_0]\,[x^{[d]\prime}-x'_0]\,[y^{[f]\prime}-y'_0]\,[y^{[d]\prime}-y'_0]}{\Sigma_1^2\Sigma_2^2}\right\}=$$

$$+N\left\{\mathsf{E}\frac{[x^{[f]\prime}-x'_0]^2}{\Sigma_1^2}\right\}\left\{\mathsf{E}\frac{[y^{[f]\prime}-y'_0]^2}{\Sigma_2^2}\right\}+$$

$$+N(N-1)\left\{\mathsf{E}\frac{[x^{[f]\prime}-x'_0]\,[x^{[d]\prime}-x'_0]}{\Sigma_1^2}\right\}\left\{\mathsf{E}\frac{[y^{[f]\prime}-y'_0]\,[y^{[d]\prime}-y'_0]}{\Sigma_2^2}\right\},$$

we have

$$\mathsf{E}[r'_{1|1}]^2=\frac{N^2}{N-1}\left\{\mathsf{E}\frac{[x^{[f]\prime}-x'_0]^2}{\Sigma_1^2}\right\}\left\{\mathsf{E}\frac{[y^{[f]\prime}-y'_0]^2}{\Sigma_2^2}\right\}$$

and finally, as

$$\mathsf{E}\frac{[x^{[f]\prime}-x'_0]^2}{\Sigma_1^2}=\frac{1}{N}\mathsf{E}\frac{\sum_{f=1}^{N}[x^{[f]\prime}-x'_0]^2}{\Sigma_1^2}=\frac{1}{N}\quad\text{and}\quad\mathsf{E}\frac{[y^{[f]\prime}-y'_0]^2}{\Sigma_2^2}=\frac{1}{N},$$

$$\mathsf{E}[r'_{1|1}]^2=\frac{1}{N-1}.$$

§ 4, 2. Although some of the expansions mentioned in text have been carried up to the order of magnitude $\left(\frac{1}{N}\right)^2$, the following collection of auxiliary formulae is confined to

those which are needed for the calculation of terms of order $\frac{1}{N}$, as otherwise they would take up too much space:

$$\mathsf{E}[dp'_{i|}]^2 = \frac{1}{N}p_{i|}(1-p_{i|}) \qquad \mathsf{E}[dp'_{i|j}]^2 = \frac{1}{N}p_{i|j}(1-p_{i|j})$$

$$\mathsf{E}[dp'_{i|j}dp'_{f|g}] = \mathsf{E}[p'_{i|j}p'_{f|g}] - p_{i|j}p_{f|g} =$$

$$= \frac{N-1}{N}p_{i|j}p_{f|g} - p_{i|j}p_{f|g} = -\frac{1}{N}p_{i|j}p_{f|g}$$

$$\mathsf{E}[dp'_{i|j}dp'_{i|g}] = -\frac{1}{N}p_{i|j}p_{i|g} \qquad \mathsf{E}[dp'_{i|}dp'_{f|}] = -\frac{1}{N}p_{i|}p_{f|}$$

$$\mathsf{E}[dp'_{i|}dp'_{i|j}] = \mathsf{E}\left\{\left[\sum_{g-1}dp'_{i|g}\right]dp'_{i|j}\right\} =$$

$$= \mathsf{E}\left\{[dp'_{i|j}]^2 + \sum_{g\neq j}dp'_{i|g}dp'_{i|j}\right\} =$$

$$= \frac{1}{N}p_{i|j}(1-p_{i|j}) - \frac{1}{N}\sum_{g\neq j}p_{i|g}p_{i|j} = \frac{1}{N}p_{i|j}(1-p_{i|})$$

$$\mathsf{E}[dp'_{|j}dp'_{i|j}] = \frac{1}{N}p_{i|j}(1-p_{|j})$$

$$\mathsf{E}[dp'_{i|}dp'_{|j}] = \mathsf{E}\left\{\left[dp'_{i|j} + \sum_{g\neq j}dp'_{i|g}\right]\left[dp'_{i|j} + \sum_{f\neq i}dp'_{f|j}\right]\right\} =$$

$$= \frac{1}{N}p_{i|j}(1-p_{i|j}) - \frac{1}{N}p_{i|j}\sum_{f\neq i}p_{f|j} - \frac{1}{N}p_{i|j}\sum_{g\neq j}p_{i|g} -$$

$$- \frac{1}{N}\left[\sum_{g\neq j}p_{i|g}\right]\left[\sum_{f\neq i}p_{f|j}\right] = \frac{1}{N}[p_{i|j} - p_{i|}p_{|j}].$$

For the proof that

$$\mathsf{E}\{[dp'_{i|j}]^h[dp'_{f|g}]^{2s-1-h}\} \quad \text{and} \quad \mathsf{E}\{[dp'_{i|j}]^h[dp'_{f|g}]^{2s-h}\}$$

contain no terms of a lower order of magnitude than $\left(\frac{1}{N}\right)^s$, I take the liberty of referring to my paper 'On the Mathematical Expectation of the Moments of Frequency-Distributions', pp. 194–200 (*Biometrika*, Vol. XII). From that it follows directly that $\mathsf{E}\{[dp'_{i|j}]^h[dp'_{i|}]^{2s-1-h}\}$, &c., likewise contain no terms of a lower order of magnitude than $\left(\frac{1}{N}\right)^s$.

Appendix

§ 4, 2, C. Noting that when the variables are mutually independent $\varphi^2 = 0$ and that all differences $p_{i|j} - p_{i|}p_{|j}$ are equal to 0, we have

$$\mathsf{E}[\varphi']^2 = \frac{1}{N}\left\{\sum_i\sum_j [1 - p_{|j}][1 - p_{i|}]\right\} + \frac{1}{N^2}\left\{\sum_i\sum_j [1 - p_{|j}]\right.$$

$$\left.[1 - p_{i|}]\right\} + \ldots = \frac{1}{N}[k-1][l-1] + \frac{1}{N^2}[k-1]$$

$$[l-1] + \ldots$$

$$\sigma^2_{[\varphi']^2} = \frac{1}{N}\left\{4\sum_i\sum_j p_{i|j} - 3\sum_i\left[\frac{1}{p^3_{i|}}(p^2_{i|}\sum_j p_{|j})^2\right] -\right.$$

$$- 3\sum_j\left[\frac{1}{p^3_{|j}}(p^2_{|j}\sum_i p_{i|})^2\right] +$$

$$\left. + 2\sum_i\sum_j\left[\frac{1}{p_{i|}p_{|j}}(p^2_{|j}\sum_f p_{f|})(p^2_{i|}\sum_g p_{|g})\right]\right\} =$$

$$= \frac{1}{N}[4 - 3 - 3 + 2] + \ldots$$

To derive the general formula for the variance of $[\varphi']^2$ it is best to start from

$$\sigma^2_{[\varphi']^2} = \mathsf{E}[\varphi']^4 - \{\mathsf{E}[\varphi']^2\}^2$$

$$[\varphi']^4 = \left\{\sum_i\sum_j \frac{n^2_{i|j}}{n_{i|}n_{|j}} - 1\right\}^2 = \left[\sum_i\sum_j \frac{n^2_{i|j}}{n_{i|}n_{|j}}\right]^2 -$$

$$- 2\sum_i\sum_j \frac{n^2_{i|j}}{n_{i|}n_{|j}} + 1 = \left[\sum_i\sum_j \frac{n^2_{i|j}}{n_{i|}n_{|j}}\right]^2 - 2[\varphi']^2 - 1$$

$$\left[\sum_i\sum_j \frac{n^2_{i|j}}{n_{i|}n_{|j}}\right]^2 = \sum_i\sum_j \frac{n^4_{i|j}}{n^2_{i|}n^2_{|j}} +$$

$$+ \sum_i\sum_j\sum_{h\neq j} \frac{n^2_{i|j}n^2_{i|h}}{n^2_{i|}n_{|j}n_{|h}} + \sum_i\sum_j\sum_{f\neq i} \frac{n^2_{i|j}n^2_{f|j}}{n_{i|}n_{f|}n^2_{|j}} +$$

$$+ \sum_i\sum_j\sum_{f\neq i}\sum_{h\neq j} \frac{n^2_{i|j}n^2_{f|h}}{n_{i|}n_{|j}n_{f|}n_{|h}}$$

The calculations are rather detailed, but offer no special difficulty.

§ 4, 3. Using the above (§ 4, 2) mentioned auxiliary formulae as a basis, the following formulae can be derived

without any special difficulties. Some of them are auxiliary formulae and some of them are of direct interest :

$$\mathsf{E}m'_{f|g} = \mathsf{E}\{\sum_i \sum_j p'_{i|j} x_f^i y_j^g\} = \sum_j \sum_i p_{i|j} x_i^f y_j^g = m_{f|g}$$

$$dm'_{f|g} = \sum_i \sum_j [p'_{i|j} - p_{i|j}] x_i^f y^g = \sum_i \sum_j dp'_{i|j} x_i^f y^g$$

$$\mathsf{E}[dm'_{f|g} dm'_{c|d}] = \frac{1}{N}[m_{f+c|g+d} - m_{f|g} m_{c|d}]$$

$$\mathsf{E}[dm'_{f|g}]^2 = \frac{1}{N}[m_{2f|2g} - m_{f|g}^2]$$

$$\mathsf{E}\mu''_{f|g} = \mu_{f|g} \qquad \mathsf{E}[d\mu''_{f|g} d\mu''_{c|d}] = \frac{1}{N}[\mu_{f+c|g+d} - \mu_{f|g}\mu_{c|d}]$$

$$\mathsf{E}[d\mu''_{f|g}]^2 = \frac{1}{N}[\mu_{2f|2g} - \mu_{f|g}^2]$$

$$\mathsf{E}[dp'_{i|j} dm'_{0|1}] = \frac{1}{N} p_{i|j}[y_j - m_{0|1}]$$

$$\mathsf{E}[dp'_{i|} dm'_{0|1}] = \frac{1}{N} p_{i|}[m_{|1}^{(i)} - m_{0|1}]$$

$$\mathsf{E}[dp'_{|j} dm'_{0|1}] = \frac{1}{N} p_{|j}[y_j - m_{0|1}]$$

$$\mathsf{E}[dp'_{i|} d\mu''_{1|1}] = \frac{1}{N}\{p_{i|}[x_i - m_{1|0}][m_{|1}^{(i)} - m_{0|1}] - p_{i|}\mu_{1|1}\}$$

$$\mathsf{E}[dp'_{i|} d\mu''_{2|0}] = \frac{1}{N}\{p_{i|}[x_i - m_{1|0}]^2 - p_{i|}\mu_{2|0}\}$$

$$\mathsf{E}[dp'_{i|} d\mu''_{0|2}] = \frac{1}{N}\{p_{i|}[m_{|1}^{(i)} - m_{0|1}]^2 + p_{i|}[\mu_{|2}^{(i)} - \mu_{0|2}]\}$$

$$\mathsf{E}m_{|1}^{(i)\prime} = \mathsf{E}\left\{\sum_j \frac{n_{i|j}}{n_{i|}} y_j\right\} = \sum_j \frac{p_{i|j}}{p_{i|}} y_j = m_{|1}^{(i)} \text{ (cf. } \textit{supra}\text{, p. 114)}$$

$$dm_{|1}^{(i)\prime} = \sum_j \left[\frac{p'_{i|j}}{p'_{i|}} - \frac{p_{i|j}}{p_{i|}}\right] y_j = \sum_j \frac{p_{i|j}}{p_{i|}}\left[\frac{dp'_{i|j}}{p_{i|j}} - \frac{dp'_{i|}}{p_{i|}} + \ldots\right] y$$

$$\mathsf{E}[dm_{|1}^{(i)\prime}]^2 = \frac{1}{N p_{i|}} \mu_{|2}^{(i)} + \ldots$$

$$\mathsf{E}[dm_{|1}^{(i)\prime} dm_{|1}^{(j)\prime}] = \frac{0}{N} + \ldots$$

$$\mathsf{E}[dp'_{i|j} dm_{|1}^{(i)\prime}] = \frac{1}{N p_{i|}} p_{i|j}[y_j - m_{|1}^{(i)}] + \ldots$$

Appendix

$$\mathsf{E}[dp'_{f|g}dm^{(i)'}_{|1}] = \frac{0}{N} + \ldots$$

$$\mathsf{E}[dp'_{i|}\,dm^{(i)'}_{|1}] = \frac{0}{N} + \ldots$$

$$\mathsf{E}[dp'_{|j}dm^{(i)'}_{|1}] = \frac{1}{Np_{i|}}p_{i|j}[y_j - m^{(i)}_{|1}] + \ldots$$

$$\mathsf{E}[dm^{(i)'}_{|1}dm'_{0|1}] = \frac{1}{N}\mu^{(i)}_{|2} + \ldots$$

$$\mathsf{E}[dm^{(i)'}_{|1}d\mu''_{2|0}] = \frac{0}{N} + \ldots$$

$$\mathsf{E}[dm^{(i)'}_{|1}d\mu''_{0|2}] = \frac{1}{N}\{\mu^{(i)}_{|3} + 2[m^{(i)}_{|1} - m_{0|1}]\mu^{(i)}_{|2}\} + \ldots =$$

$$= \frac{1}{N}\left\{\frac{1}{p_{i|}}\sum_j p_{i|j}[y_j - m_{0|1}]^3 - [m^{(i)}_{|1} - \right.$$

$$\left. - m_{0|1}]\mu^{(i)}_{|2} - [m^{(i)}_{|1} - m_{0|1}]^3\right\} + \ldots$$

$$\mathsf{E}[dm^{(i)'}_{|1}d\mu''_{1|1}] = \frac{1}{Np_{i|}}\{\sum_j p_{i|j}[x_i - m_{1|0}]\,[y_j - m_{0|1}]^2 -$$

$$- [m^{(i)}_{|1} - m_{0|1}]\sum_j p_{i|j}[x_i - m_{1|0}]\,[y_j - m_{0|1}]\} + \ldots =$$

$$= \frac{1}{N}\{[x_i - m_{1|0}]\sum_j p^{(i)}_{|j}[y_j - m_{0|1}]^2 - [x_i - m_{1|0}]$$

$$[m^{(i)}_{|1} - m_{0|1}]^2\} + \ldots = \frac{1}{N}[x_i - m_{1|0}]\mu^{(i)}_{|2} + \ldots$$

$$\mathsf{E}\mu'_{f|g} = \mu_{f|g} - \frac{1}{N}\{(f+g)\mu_{f|g} -$$

$$- \frac{1}{2}f(f-1)\mu_{f-2|g}\mu_{2|0} - \frac{1}{2}g(g-1)\mu_{f|g-2}\mu_{0|2} -$$

$$- fg\mu_{f-1|g-1}\mu_{1|1}\} + \ldots$$

$$\mathsf{E}[d\mu'_{f|g}]^2 = \frac{1}{N}\{\mu_{2f|2g} - \mu^2_{f|g} + f^2\mu^2_{f-1|g}\mu_{2|0} + g^2\mu^2_{f|g-1}\mu_{0|2} -$$

$$- 2f\mu_{f+1|g}\mu_{f-1|g} - 2g\mu_{f|g+1}\mu_{f|g-1} + 2fg\mu_{f-1|g}\mu_{f|g-1}\mu_{1|1}\} + \ldots$$

§ 4, 3, B and C. Noting that for linear regression of Y and X

$$m_{|1}^{(i)} - m_{0|1} = r_{1|1}\sqrt{\frac{\mu_{0|2}}{\mu_{2|0}}}[x_i - m_{1|0}],$$

we have

$$\Sigma_i p_{i|}[x_i - m_{1|0}][m_{|1}^{(i)} - m_{0|1}]^2 = \frac{\mu_{0|2}}{\mu_{2|0}} r_{1|1}^2 \mu_{3|0} = \mu_{0|2}\sqrt{\mu_{2|0}} r_{1|1}^2 r_{3|0}$$

$$\Sigma_i p_{i|}[m_{|1}^{(i)} - m_{0|1}]^4 = \mu_{0|2}^2 r_{1|1}^4 r_{4|0}$$

$$\Sigma_i p_{i|}[x_i - m_{1|0}]^2[m_{|1}^{(i)} - m_{0|1}]^2 = \mu_{0|2}\mu_{2|0} r_{1|1}^2 r_{4|0}$$

$$\Sigma_i p_{i|}[x_i - m_{1|0}][m_{|1}^{(i)} - m_{0|1}]^3 = \sqrt{\mu_{2|0}\mu_{0|2}^3}\, r_{1|1}^3 r_{4|0}.$$

From the identity

$$\mu_{2|2} = \Sigma_i \Sigma_j p_{ij}[x_i - m_{1|0}]^2[y_j - m_{0|1}]^2 =$$

$$= \Sigma_i \{p_{i|}[x_i - m_{1|0}]^2 \Sigma_j p_{|j}^{(i)}[y_j - m_{0|1}]^2\} =$$

$$= \Sigma_i \{p_{i|}[x_i - m_{1|0}]^2 (\Sigma_j p_{|j}^{(i)}[y_j - m_{|1}^{(i)}]^2 + [m_{|1}^{(i)} - m_{0|1}]^2)\} =$$

$$= \Sigma_i p_{i|}[x_i - m_{1|0}]^2 \mu_{|2}^{(i)} + \Sigma_i p_{i|}[x_i - m_{1|0}]^2[m_{|1}^{(i)} - m_{0|1}]^2$$

it is easily seen that if the regression of Y on X is linear,

$$\Sigma_i p_{i|}[x_i - m_{1|0}]^2 \mu_{|2}^{(i)} = u_{2|0}\mu_{0|2}[r_{2|2} - r_{1|1}^2 r_{4|0}]$$

$$\Sigma_i p_{i|}[m_{|1}^{(i)} - m_{0|1}]^2 \mu_{|2}^{(i)} = \frac{\mu_{0|2}}{\mu_{2|0}} r_{1|1}^2 \Sigma_i p_{i|}[x_i - m_{1|0}]^2 \mu_{|2}^{(i)} =$$

$$= \mu_{0|2}^2 r_{1|1}^2 [r_{2|2} - r_{1|1}^2 r_{4|0}]$$

$$\Sigma_i p_{i|}[x_i - m_{1|0}][m_{|1}^{(i)} - m_{0|1}]\mu_{|2}^{(i)} = \sqrt{\mu_{2|0}\mu_{0|2}^3}\, r_{1|1}[r_{2|2} - r_{1|1}^2 r_{4|0}].$$

Similarly from the identity

$$\mu_{1|3} = \Sigma_i p_{i|}[x_i - m_{1|0}]\mu_{|3}^{(i)} + 3\Sigma_i p_{i|}[x_i - m_{1|0}][m_{|1}^{(i)} - m_{0|1}]\mu_{|2}^{(i)} +$$

$$+ \Sigma_i p_{i|}[x_i - m_{1|0}]\,[m_{|1}^{(i)} - m_{0|1}]^3$$

on the assumption that the regression of Y on X is linear, we have

$$\Sigma_i p_{i|}[x_i - m_{1|0}]\mu_{|3}^{(i)} = \sqrt{\mu_{2|0}\mu_{0|2}^3}[r_{1|3} - 3r_{1|1}r_{2|2} + 2r_{1|1}^3 r_{4|0}]$$

$$\Sigma_i p_{i|}[m_{|1}^{(i)} - m_{0|1}]\mu_{|3}^{(i)} = \mu_{0|2}^2 r_{1|1}[r_{1|3} - 3r_{1|1}r_{2|2} + 2r_{1|1}^3 r_{4|0}],$$

and from the identity

$$\mu_{1|2} = \Sigma_i p_{i|}[x_i - m_{1|0}]\mu_{|2}^{(i)} + \Sigma_i p_{i|}[x_i - m_{1|0}]\,[m_{|1}^{(i)} - m_{0|1}]^2,$$

when the regression of Y on X is linear,

$$\Sigma_i p_{i|}[x_i - m_{1|0}]\mu_{|2}^{(i)} = \mu_{0|2}\sqrt{\mu_{2|0}}[r_{1|2} - r_{1|1}^2 r_{3|0}].$$

Appendix

From $\eta^2_{y|x} = 1 - \frac{1}{\mu_{0|2}} \Sigma_i p_{i|} \mu^{(i)}_{|2}$ we obtain

$$\Sigma_i p_{i|} \mu^{(i)}_{|2} = \mu_{0|2}[1 - \eta^2_{y|x}].$$

If Y is homoscedastically connected with X, we have $\mu^{(i)}_{|2} = \mu_{0|2}[1 - \eta^2_{y|x}]$. When the regression of Y on X is linear $\eta^2_{y|x} = r^2_{1|1}$, and hence $\Sigma_i p_{i|} \mu^{(i)}_{|2} = \mu_{0|2}[1 - r^2_{1|1}]$; if also Y is homoscedastically connected with X, $\mu^{(i)}_{|2} = \mu_{0|2}[1 - r^2_{1|1}]$.

Putting $\mu^{(i)}_{|2} = \mu_{0|2}[1 - r^2_{1|1}]$ in

$$\Sigma_i p_{i|}[x_i - m_{1|0}]^2 \mu^{(i)}_{|2} = \mu_{2|0}\mu_{0|2}[r_{2|2} - r^2_{1|1} r_{4|0}]$$

and in $\Sigma_i p_{i|}[x_i - m_{1|0}] \mu^{(i)}_{|2} = \mu_{0|2} \sqrt{\mu_{2|0}} [r_{1|2} - r^2_{1|1} r_{3|0}]$,

for the case when the regression of Y on X is linear and the connexion of Y with X is homoscedastic, we have

$$1 - r^2_{1|1} = r_{2|2} - r^2_{1|1} r_{4|0}$$
$$0 = r_{1|2} - r^2_{1|1} r_{3|0}.$$

§ 4, 3, D. If the values $\mathsf{E}u'$, $\mathsf{E}[u']^2$, $\sigma_{u'}$, $\sigma_{[u']^2}$, are computed at the same time, it is not necessary to carry out all the computations four times. The work can be considerably facilitated, as follows :

If U' is expressed $U' = c + d^I + d^{II} + \ldots$ where c is the sum of terms, which do not contain the differences $dp'_{i|}$, &c., d^I the sum of terms which contain only the first powers of the differences $dp'_{i|}$, &c., we have

$$[u']^2 = c^2 + 2cd^I + \{[d^I]^2 + 2cd^{II}\} + \ldots$$
$$[u']^4 = c^4 + 2c^3 d^I + \{6c^2[d^I]^2 + 4c^3 d^{II}\} + \ldots$$

Hence

$$\mathsf{E}u' = c + \mathsf{E}d^{II} + \ldots$$
$$u' - \mathsf{E}u' = d^I + \{d^{II} - \mathsf{E}d^{II}\} + \ldots$$
$$\mathsf{E}[u']^2 = c^2 + \{\mathsf{E}[d^I]^2 + 2c\mathsf{E}d^{II}\} + \ldots$$
$$\mathsf{E}[u']^4 = c^4 + \{6c^2\mathsf{E}[d^I]^2 + 4c^3\mathsf{E}d^{II}\} + \ldots$$
$$\sigma^2_{u'} = \mathsf{E}[u' - \mathsf{E}u']^2 = \mathsf{E}[d^I]^2 + \ldots$$
$$\sigma^2_{[u']^2} = \mathsf{E}\{[u']^2 - \mathsf{E}[u']^2\}^2 = \mathsf{E}[u']^4 - \{\mathsf{E}[u']^2\}^2 =$$
$$= 4c^2\mathsf{E}[d^I]^2 + \ldots$$

Thus the calculation of $\mathsf{E}[u']^2$ and of $\sigma_{[u']^2}$ requires, if we are content with the approximation up to the terms of order $\frac{1}{N}$, only the knowledge of the values $\mathsf{E}[d^I]^2$ and of $\mathsf{E}d^{II}$, which are necessary for the computation of $\mathsf{E}u'$ and of $\sigma_{u'}$.

Similarly the calculation of the variances

$$[\zeta'_{y|x}]^2 = [\eta'_{y|x}]^2 - [r'_{1|1}]^2$$

can be simplified. If, for shortness, we write η' instead of $\eta'_{y|x}$ and r' instead of $r'_{1|1}$ and start from expansion analogous to those above

$$r' = c + d^I + d^{II} + \ldots$$
$$[\eta']^2 = k + \Delta^I + \Delta^{II} + \ldots,$$

we have

$$[r']^2 = c^2 + 2cd^I + \{[d^I]^2 + 2cd^{II}\} + \ldots$$
$$[\eta']^2[r']^2 = kc^2 + \{c^2\Delta^I + 2kcd^I\} + \\ + \{c^2\Delta^{II} + 2c\Delta^I d^I + k[d^I]^2 + 2kcd^{II}\} + \ldots$$
$$\mathsf{E}\{[\eta']^2 - \mathsf{E}[\eta']^2\}\{[r']^2 - \mathsf{E}[r']^2\} = \mathsf{E}\{[\eta']^2[r']^2\} - \\ - \{\mathsf{E}[\eta']^2\}\{\mathsf{E}[r']^2\} = 2c\mathsf{E}[\Delta^I d^I] + \ldots$$
$$\sigma^2_{[\zeta']^2} = \mathsf{E}\{([\eta']^2 - \mathsf{E}[\eta']^2) - ([r']^2 - \mathsf{E}[r']^2)\}^2 = \\ = \sigma^2_{[\eta']^2} + \sigma^2_{[r']^2} - 2\mathsf{E}\{[\eta']^2 - \mathsf{E}[\eta']^2\}\{[r']^2 - \mathsf{E}[r']^2\} = \\ = \sigma^2_{[\eta']^2} + \sigma^2_{[r']^2} - 4c\mathsf{E}[\Delta^I d^I] - \ldots$$

§ 4, 4. Introducing the abbreviated notation $\mathsf{E}z' = m_{1|0}$, $\mathsf{E}z'w' = m_{1|1}$ where $\frac{1}{c}\mathsf{E}z' - \frac{1}{c^2}\mathsf{E}z'(w' - c) = \frac{2m_{1|0}}{c} - \frac{m_{1|1}}{c^2}$, then the condition

$$\frac{\partial\left[\frac{2m_{1|0}}{c} - \frac{m_{1|1}}{c^2}\right]}{\partial c} = 2\left[\frac{m_{1|1}}{c^3} - \frac{m_{1|0}}{c^2}\right] = 0$$

yields the value $c = \frac{m_{1|1}}{m_{1|0}}$, after the substitution of which we find $\frac{2m_{1|0}}{c} - \frac{m_{1|1}}{c^2} = \frac{m^2_{1|0}}{m_{1|1}}$.

Further, putting $\mathsf{E}w' = m_{0|1}$, $\mathsf{E}[w']^2 = m_{0|2}$, where

$$\frac{1}{c}\mathsf{E}z' - \frac{1}{c^2}\mathsf{E}z'(w'-c) + \frac{1}{c^2}\mathsf{E}(w'-c)^2 = \frac{2m_{1|0}}{c} - \frac{m_{1|1}}{c^2} + \frac{m_{0|2}}{c^2} - \frac{2m_{0|1}}{c} + 1$$

we obtain similarly

$$c = \frac{m_{0|1} - m_{1|0}}{m_{0|2} - m_{1|1}}.$$

§ 4, 5. The formulae for $\mu_{2|0}$, $\mu_{0|2}$, $r_{1|1}$, when both the variables can assume only two values each, have been derived above (Chap. IV, § 6). Since

$$p^3 + (1-p)^3 = 1 - 3p(1-p)$$

we have

$$\mu_{4|0} = [p_{1|}p_{2|}^4 + p_{2|}p_{1|}^4][x_1 - x_2]^4 = p_{1|}p_{2|}[1 - 3p_{1|}p_{2|}][x_1 - x_2]^4$$

$$r_{4|0} = \frac{\mu_{4|0}}{\mu_{2|0}^2} = \frac{1}{p_{1|}p_{2|}} - 3.$$

Substituting $p_{1|1} = \delta + p_{1|}p_{|1}$, $p_{1|2} = -\delta + p_{1|}p_{|2}$, &c., we obtain

$$\mu_{3|1} = [x_1 - x_2]^3[y_1 - y_2][p_{1|1}p_{2|}^3p_{|2} - p_{1|2}p_{2|}^3p_{|1} - p_{2|1}p_{1|}^3p_{|2} + p_{2|2}p_{1|}^3p_{|1}] = [x_1 - x_2]^3[y_1 - y_2]\{\delta[p_{2|}^3p_{|2} - p_{2|}^3p_{|1} + p_{1|}^3p_{|2} + p_{1|}^3p_{|1}] + p_{1|}p_{2|}p_{|1}p_{|2}[p_{2|} - p_{2|} - p_{1|} + p_{1|}]\} = [x_1 - x_2]^3[y_1 - y_2]\delta[p_{1|}^3 + p_{2|}^3]$$

$$r_{3|1} = r_{1|1}r_{4|0}$$

$$\mu_{2|2} = [x_1 - x_2]^2[y_1 - y_2]^2[p_{1|1}p_{2|}^2p_{|2}^2 + p_{1|2}p_{2|}^2p_{|1}^2 + p_{2|1}p_{1|}^2p_{|2}^2 + p_{2|2}p_{1|}^2p_{|1}^2] = [x_1 - x_2]^2[y_1 - y_2]^2\{\delta[p_{1|} - p_{2|}][p_{|1} - p_{|2}] + p_{1|}p_{2|}p_{|1}p_{|2}\}$$

$$r_{2|2} = 1 + \frac{[p_{1|} - p_{2|}][p_{|1} - p_{|2}]\delta}{p_{1|}p_{2|}p_{|1}p_{|2}}.$$

Since $r_{3|1} = r_{1|1}r_{4|0}$ and $r_{1|3} = r_{1|1}r_{0|4}$, it is only necessary to put the above values of the parameters r in the formulae which have been derived for mutually linear regression in order to obtain the mathematical expectation and the variance of $r'_{1|1}$.

Appendix

CHAPTER VII

§ 1. For the scheme of draws without replacement we have found above (Chap. VI, § 2, 1):

$$\mathsf{E}[dp'_{i|j}]^2 = \frac{A-N}{A-1}\frac{1}{N}p_{i|j}(1-p_{i|j})$$

$$\mathsf{E}[dp'_{i|}]^2 = \frac{A-N}{A-1}\frac{1}{N}p_{i|}(1-p_{i|})$$

$$\mathsf{E}[dp'_{|j}]^2 = \frac{A-N}{A-1}\frac{1}{N}p_{|j}(1-p_{|j}).$$

Proceeding from

$$\mathsf{E}n_{i|j}n_{f|g} = \sum_h P_h h(N-h)\frac{p_{f|g}}{1-p_{i|j}} = \frac{p_{f|g}}{1-p_{i|j}}\{N\sum_h P_h h -$$

$$-\sum_h P_h h^2\} = \frac{p_{f|j}}{1-p_{i|j}}\left\{N^2 p_{i|j} - N^2 p_{i|j}\frac{B-1}{A-1} -\right.$$

$$\left. - Np_{i|}\left[1-\frac{B-1}{A-1}\right]\right\} = \frac{A}{A-1}N(N-1)p_{i|j}p_{f|}$$

we obtain further

$$\mathsf{E}[dp'_{i|j}dp'_{f|g}] = \frac{1}{N^2}\{\mathsf{E}[n_{i|j}n_{f|g}] - [\mathsf{E}n_{i|j}]\,[\mathsf{E}n_{f|g}]\} =$$

$$= -\frac{A-N}{A-1}\frac{1}{N}p_{i|j}p_{|g}$$

and hence

$$\mathsf{E}[dp'_{i|}\,dp'_{f|}] = -\frac{A-N}{A-1}\frac{1}{N}p_{i|}p_{f|}$$

$$\mathsf{E}[dp'_{i|}dp'_{i|j}] = \frac{A-N}{A-1}\frac{1}{N}p_{i|j}(1-p_{i|})$$

$$\mathsf{E}[dp'_{i|}dp'_{|j}] = \frac{A-N}{A-1}\frac{1}{N}[p_{i|j}-p_{i|}p_{|j}].$$

Hence, the mathematical expectations of second powers of differences $dp'_{i|j}$, &c., become equal in the case of draws without replacement, to $\frac{A-N}{A-1}$ times the corresponding mathematical expectations in the case of independent trials. Thus, if we proceed from the expansion of a function of

the empirical values in the form (cf. Chap. VI, § 4, 3, D)

$$u' = c + d^I + d^{II} + \ldots$$

and denote by $\frac{1}{N}D^{II}$ the mathematical expectation of d^{II}, in the case of independent trials we obtain for the case of draws without replacement

$$\mathsf{E}u' = c + \frac{A-N}{A-1}\frac{1}{N}D^{II} + \ldots$$

Similarly, it may be seen that in the case of draws with addition $\mathsf{E}[dp'_{i|j}]^2 = \frac{A+N}{A+1}\frac{1}{N}p_{i|j}(1-p_{i|j})$, &c., and hence

$$\mathsf{E}u' = c + \frac{A+N}{A+1}\frac{1}{N}D^{II} + \ldots$$

In this case one must not overlook the fact that $\frac{A+N}{A+1} > 1$ and that for small A $\frac{A+N}{A+1}$ can be of the order of magnitude of N. Accordingly the following terms of the series can be of the same order of magnitude, even for large N, so that the approximation reached in this way will become illusory.

§ 2. As

$$\mu'_{0|2} = \sum_j p'_{|j}[y_j - m'_{0|1}]^2 = \sum_j p'_{|j}[y_j - m_{0|1}]^2 -$$

$$- 2\sum_j p'_{|j}[m'_{0|1} - m_{0|1}]\,[y_j - m_{0|1}] + [m'_{0|1} - m_{0|1}]^2 = \mu''_{0|2} -$$

$$- [dm'_{0|1}]^2 \qquad \mathsf{E}\mu''_{0|2} = \mu_{0|2}$$

in all three cases, and

$\mathsf{E}[dm'_{0|1}]^2 = \frac{1}{N}\mu_{0|2}$ in the case of draws with replacement,

$\mathsf{E}[dm'_{0|1}]^2 = \frac{A-N}{A-1}\frac{1}{N}\mu_{0|2}$ in the case of draws without replacement, $\mathsf{E}[dm'_{0|1}]^2 = \frac{A+N}{A+1}\frac{1}{N}\mu_{0|2}$ in the case of draws with addition,

we obtain:

$$\mathsf{E}\mu'_{0|2} = \mu_{0|2} - \frac{1}{N}\mu_{0|2} = \frac{N-1}{N}\mu_{0|2}$$

in the case of draws with replacement,

$$\mathsf{E}\mu'_{0|2} = \mu_{0|2} - \frac{A-N}{A-1}\frac{1}{N}\mu_{0|2} = \frac{A}{A-1}\frac{N-1}{N}\mu_{0|2}$$

in the case of draws without replacement,

$$\mathsf{E}\mu'_{0|2} = \mu_{0|2} - \frac{A-N}{A+1}\frac{1}{N}\mu_{0|2} = \frac{A}{A+1}\frac{N-1}{N}\mu_{0|2}$$

in the case of draws with addition.

§ 3, 1. As

$$\sum_{f=1}^{N}\{x^{[f]\prime} - x'_0\}\{y^{[f]\prime} - y'_0\}\sum_{f=1}^{N} = \{x^{[f]\prime} - m_{1|0}\}\{y^{[f]\prime} - m_{1|0}\} - \\ - N[x'_0 - m_{1|0}][y'_0 - m_{0|1}]$$

and, again, for mutually independent trials,

$$\mathsf{E}[x'_0 - m_{1|0}][y'_0 - m_{0|1}] = \frac{1}{N^2}\mathsf{E}\Big\{\sum_{f=1}^{N}[x^{[f]\prime} - m_{1|0}]\Big\}$$

$$\Big\{\sum_{f=1}^{N}[y^{[f]\prime} - m_{0|1}]\Big\} = \frac{1}{N^2}\mathsf{E}\Big\{\sum_{f=1}^{N}[x^{[f]\prime} - m_{1|0}][y^{[f]\prime} - m_{0|1}] +$$

$$+\sum_{f=1}^{N}\sum_{g\neq f}[x^{[f]\prime} - m_{1|0}][y^{[g]\prime} - m_{0|1}]\Big\} = \frac{1}{N}\mu_{1|1},$$

it follows that

$$\mathsf{E}\Big\{\sum_{f=1}^{N}[x^{[f]\prime} - x'_0][y^{[f]\prime} - y'_0]\Big\} = N\mu_{1|1} - \mu_{1|1} = (N-1)\mu_{1|1}.$$

§ 3, 2. As the trials of each series are independent, we have

$$\mathsf{E}[x_0^{[h]\prime} - m_{1|0}][y_0^{[h]\prime} - m_{0|1}] = \frac{1}{n}\mu_{1|1}.$$

As again x'_0 is the arithmetic mean of the values $x_0^{[h]\prime}$, and y'_0 the arithmetic mean of the values $y_0^{[h]\prime}$, we have from the above general formula (§ 3, 1)

$$\mathsf{E}\Big\{\sum_{h=1}^{r}[x_0^{[h]\prime} - x'_0][y_0^{[h]\prime} - y'_0]\Big\} = (r-1)\frac{1}{n}\mu_{1|1}.$$

Appendix

§ 3, 3. Denoting the numerator of Q by Z and the denominator of Q by T, it can be easily shown that in the case of normal stability for any value of $k \mathsf{E}ZT^k = \mathsf{E}T^{k+1}$ and obviously for $k = -1$, $\mathsf{E}\frac{Z}{T} = \mathsf{E}Q = 1$. Proceeding from

$$\sum_{f=1}^{r\,n}[x^{[f]\prime} - x'_0][y^{[f]\prime} - y'_0] = \sum_{f=1}^{r\,n} x^{[f]\prime}y^{[f]\prime} - rnx'_0y'_0$$

and noting that the law of dependence remains constant and all the trials are mutually independent, we obtain

$$\mathsf{E}T^{k+1} = \mathsf{E}TT^k =$$

$$= \frac{1}{rn-1}\Big\{\mathsf{E}\Big[\sum_{f=1}^{r\,n} x^{[f]\prime}y^{[f]\prime}\Big]T^k - \frac{1}{rn}\mathsf{E}\Big[\sum_{f=1}^{r\,n} x^{[f]\prime}\Big]$$

$$\Big[\sum_{f=1}^{r\,n} y^{[f]\prime}\Big]T^k\Big\} = \frac{1}{rn-1}\Big\{rn\mathsf{E}x^{[f]\prime}y^{[f]\prime}T^k - \frac{1}{rn}$$

$$\mathsf{E}\Big[\sum_{f=1}^{r\,n} x^{[f]\prime}y^{[f]\prime} + \sum_{f-1}^{r\,n}\sum_{g\neq f} x^{[f]\prime}y^{[g]\prime}\Big]T^k\Big\} =$$

$$= \mathsf{E}x^{[f]\prime}y^{[f]\prime}T^k - \mathsf{E}x^{[f]\prime}y^{[g]\prime}T^k.$$

In the same way it may be shown that

$$\mathsf{E}ZT^k = \mathsf{E}x^{[f]\prime}y^{[f]\prime}T^k - \mathsf{E}x^{[f]\prime}y^{[g]\prime}T^k$$

and hence $$\mathsf{E}ZT^k = \mathsf{E}T^{k+1}.$$

NOTES AND BIBLIOGRAPHY *

THE present treatise is not intended to serve as an introduction to the practical application of statistical methods. For this latter purpose A. Tschuprow recommended in the first place G. Udny Yule's *An Introduction to the Theory of Statistics.* This famous textbook, revised by G. Udny Yule and M. G. Kendall, is now in its eleventh edition (London: Charles Griffin Company Ltd.; 1937). Tschuprow also mentioned Truman L. Kelley's *Statistical Methods* (New York: Macmillan; 1923) and pointed out that American statistical literature is notably devoted to the simplification and improvement of methods of calculation of correlation coefficients, ratios, &c., which may be followed in the *Journal of American Association.*

The translator would like to suggest in addition to the above-mentioned books R. A. Fisher's *Statistical Methods for Research Workers* (7th edition; Edinburgh and London: Oliver & Boyd; 1938), the first edition of which appeared in the same year as the German original of the present book. Fisher's statistical methods may be considered as a most practical guide to the application of statistical methods, particularly in the field of biology. Arthur L. Bowley's *Elements of Statistics* (6th edition; London: P. S. King & Son, Ltd.; 1937) will be of particular use to investigators in the field of social and economic science.

Those readers who are sufficiently equipped with mathematical knowledge to understand such articles as, for example, L. Isserlis' 'On the Partial Correlation Ratio' (*Biometrika*, Vol. X, 1914, pp. 391–411, and Vol. XII, 1916–17, pp. 50–66), will find ample bibliography on correlation, apart from the articles referred to below by A. Tschuprow (cf. pp. 184 *et seq.*), in Yule-Kendall's *Introduction*, pp. 509 *et seq.*, as well as in the articles on 'Recent

* The translator thought it expedient to substitute a survey of contemporary English literature for the author's introductory notes. However, bibliography relating to separate chapters has been translated from the original.

Advances in Mathematical Statistics ', edited or written by J. O. Irwin and published periodically in the *J. Roy. Stat. Soc.* (the last bibliography appeared in Vol. CI, Pt. II, pp. 394 *et seq.*; 1938). Further, we should like to mention some useful statistical tables which may considerably facilitate the computer's task. The most important of these are Karl Pearson's *Tables for Statisticians and Biometricians* (two parts; Cambridge: University Press); *The Kelley Statistical Tables*, by Truman L. Kelley (New York: The Macmillan Company; 1938), and *Statistical Tables for Biological Agricultural and Medical Research*, by R. A. Fisher and F. Yates (Edinburgh and London: Oliver & Boyd; 1938).

There are also some special tables for the purpose of correlation, as, for instance, T. L. Kelley, *Tables to Facilitate the Calculation of Partial Coefficients of Correlation and Regression Equations* (Bulletin of the University of Texas, No. 27; 1916), J. R. Miner, *Tables of* $\sqrt{1-r^2}$ and $1-r^2$ *for Use in Partial Correlation* (Baltimore: The Johns Hopkins' Press; 1922), and F. N. David, *Tables of the Ordinates and Probability Integral of the Distribution of the Correlation Coefficient in Small Samples* (issued by the Biometrika Office, University College, London, and printed at the University Press, Cambridge, England; 1938). Also a small elementary textbook by C. B. Davenport and Merle P. Ekas on *Statistical Methods in Biology, Medicine, and Psychology* (4th edition, completely revised; New York: John Wiley & Sons; 1936), as well as the eleventh edition of Yule's *Introduction*, contain very useful tables.

L. Isserlis has translated into English Tschuprow's paper on ' The Mathematical Theory of the Statistical Methods Employed in the Study of Correlation in the Case of Three Variables' (*Trans. Camb. Phil. Soc.*, Vol. 23, 1928, p. 337).

M. K.

CHAPTER I

From about the middle and during the latter half of the nineteenth century, statisticians were engaged upon the discovery of those statistical methods which nowadays are considered elementary; the French statisticians Guerry and Dufau deserve particular mention. These methods are very popular in Russian statistical textbooks, where, until recently, they were understood

as modalities of inductive method of competitive variation referred to by J. St. Mill's Logic. In order to give an example of a more recent date, an article by F. Toennies, ' Eine Neue Methode der Vergleichung Statistischer Reihen ' (*Jahrbuch für Gesetzgebung, Verwaltung, und Volkswirtschaft*, Vol. 33, 1909) may be mentioned.

§ 2, B. In a similar way to the above, the correlation coefficient is introduced with reference to Fechner's Index-Numbers by L. March in his inspiring treatise on ' Comparison Numérique de Courbes Statistiques ' (*Journal de la Société de Statistique de Paris*, 1905).

§ 2, C. The coefficient ϱ was introduced by the psychologist C. Spearman. The collocation of the problem of rank-correlation in the system of the theory of correlation and the proof that in the case of normal correlation the correlation coefficient r is connected with Spearman's coefficient ϱ by the formula $r = 2 \sin\left(\frac{\pi}{6}\varrho\right)$ is due to K. Pearson (vide K. Pearson, *On Further Methods of Determining Correlation*, Drapers' Company Research Memoirs, Biometric Series, IV, 1907).

CHAPTER II

The conception of the subject and the task of statistical correlation inquiry referred to in this chapter is represented in more detailed form in my *Theory of Statistics* (Russian, 2nd edition, St. Petersburg, 1910); vide A. Tschuprow, ' Die Aufgaben der Theorie der Statistik ' (*Jahrbuch für Gesetzgebung, Verwaltung, und Volkswirtschaft*, 1905) and A. Kaufmann, *Theorie und Methoden der Statistik*, Part I, Chap. V (Tübingen, 1913).

CHAPTER III

§ 4. The presentation in the text corresponds in *a priori* conception to the ' generalized idea of correlation ' which K. Pearson has developed in his treatise *On the General Theory of Skew Correlation and Non-Linear Regression* (Drapers' Company Research Memoirs, Biometric Series, II, 1905). With reference to Fr. Galton, K. Pearson originally stated the notion of being correlated defined above (§ 4, 3). Vide K. Pearson, ' Regression,

Hereditary and Panmixia' (*Phil. Trans.*, A, Vol. 187, pp. 256–7; 1897). For detailed bibliography, vide A. Tschuprow, 'Ziele und Wege der Stochastischen Grundlegung der Statistischen Theorie', § 2, 3 (*Nordisk Statistisk Tidskrift*, Vol. III; 1924).

§ 5. J. Kleiber appears to be the first person to draw attention to the fact that when working out measurements of functionally connected magnitudes, which are liable to chance errors, one obtains incompatible equations if at one time one considers X and at another time Y as independent variables. ('Über den Abrundungsfehler Meteorologischer Zahlen', *Meteor. Zeitschrift*, Vol. V; 1888); however, Kleiber does not take into consideration the nature of this discrepancy and tries to attribute it to the smoothing-error. Sresnewski has proved, however, that this is not the case ('Über Abrundungsfehler', *Meteor. Zeitschrift*, Vol. VI; 1889). This problem was put into a more precise conception and solved under certain restricted conditions by Karl Pearson ('On Lines and Planes of Closest Fit to Systems of Points in Space', *Phil. Mag.*, Ser. 6, Vol. II; 1901). This problem has been treated in a similar way by various authors with partial reference to K. Pearson and partial independence of him; vide, for instance, the treatises by C. Gini, *Sull'interpolazione di una retta quando i valori della variable indipendente sono affetti da errori accidentali*, and L. J. Reed, 'Fitting Straight Lines', in the second volume of *Metron* (1921). W. Wirth takes an exceptional stand, according to which the point is not the finding out of the law of functional relationship between non-chance variables, but (to clothe his thoughts in my expressions) the point in question is the method of determining the stochastic connexion between two chance variables (vide W. Wirth, 'Spezielle Psycho-Physische Massmethoden' in *Handbuch der Biologischen Arbeitsmethoden*, edited by E. Abderhalden, Section 6A, number 1; Berlin, Vienna; 1920; and the articles referring to this treatise by E. Czuber and W. Wirth in *Archiv für die Gesamte Psychologie*, Vols. XLI (1921) and XLIV (1923); particularly the article by W. Wirth on 'K. Pearson's Angepasste Grade (Best Fitting Straight Line) und die Mittlere Regression', (*Archiv für die Gesamte Psychologie*, Vol. XLIV).

K. Pearson very instructively elucidates in his 'Notes on the History of Correlation' (*Biometrika*, Vol. XIII) the distinction between the statistical correlation meaning and the standpoint

of non-statisticians on the hand of Gauss'-Bravais'-Galton's contributions to the theory of correlation.

CHAPTER IV

The presentation rests upon my treatise in the second volume of *Investigations by Russian Scientists Abroad* (Russian, Berlin, 1923). The problems treated in Chapters I and V are usually handled monographically in connexion with those questions which are attached to the frame of this systematic presentation of Chapter VI. In consideration of this it appears to be more expedient to concentrate the corresponding bibliography upon Chapter VI.

§ 2, 1. (Cf. Chap. V, § 6.) W. F. Sheppard has suggested to compare the differences $p'_{i|j} - p'_{j|}p'_{|j}$ with their probable errors in order to determine the absence of mutual independence between the variables, whereby Sheppard characterizes this method as ' an extension of the ordinary method (used largely by Prof. Lexis and Prof. Edgeworth) for testing the stability of statistical ratios ' (' On the Application of the Theory of Error to Cases of Normal Distribution and Normal Correlation ', § 21 ; *Phil. Trans.*, A, Vol. 192, p. 130 ; 1899). Sheppard stops at the consideration of totality of individual values, without having constructed any comprehensive coefficient. The introduction of the coefficient φ (mean square contingency) is due to K. Pearson, *On the Theory of Contingency and its Relation to Association and Normal Correlation* (Drapers' Company Research Memoirs, Biometric Series, I ; 1904).

§ 2, 1, and § 7. (Cf. Chap. V, § 7.) The value of tetrachoric correlation coefficient is computed in the same way as above, however, under assumption that the variables X and Y can obtain only the values 1 and 0 ; cf. G. U. Yule, ' On the Methods of Measuring Association Between two Attributes ', pp. 596, 606 *et seq.* (*J. Roy. Stat. Soc.*, Vol. 75 ; 1912) ; L. von Bortkiewicz, Review of Charlier's textbook, pp. 346–7 (*Nordisk Statistisk Tidskrift*, Vol. I, 1922) ; vide further, C. Gini, ' Indici di omofilia e loro relazioni col coefficiente di correlazione e con gli indici di attrazione ', p. 600, and note 1 to p. 602 (*Atti del Reale Istituto Veneto*, t. 74, Parte seconda, Venezia, 1915). C. V. Charlier (*Vorlesungen über die Grundzüge der Mathe-*

matischen Statistik, pp. 105–14 ; 1920) obtains the same expression for the correlation coefficient indirectly by the equation of correlation surfaces. S. D. Wicksell (' Some Theorems in the Theory of Probability, with Special Reference to Their Importance in the Theory of Homograde Correlation ', *Svenska Aktuarieföreningens Tidskrift*, 1916) arrives at it by the regression equations.

The consideration treated in the text which starts from the assumption that both the variations can obtain only two different values each is to be distinguished from the presentation of a problem based on a supposition that any quantity of possible values of variable can be gathered into two groups each, in which it would appear to be the task of theoretical treatment of the problem, to show in what way the coefficients characterizing the law of dependence can be obtained on the basis of such a tetrachoric table. This is the problem presented by Karl Pearson in his consideration of the tetrachoric table. In order that this task be solvable the assumption of a definite law of dependence must be added ; the normal correlation comes into consideration in the first place as a matter of course ; vide K. Pearson ' On the Correlation of Characters Not Quantitatively Measurable ' (*Phil. Trans.*, A, Vol. 1905 ; 1901) ; K. Pearson, *On a Novel Method of Regarding the Association of Two Variates Classed Solely in Alternate Categories* (Drapers' Company Research Memoirs, Biometric Series, VII ; 1912) ; K. Pearson and D. Heron, ' On Theories of Association ' (*Biometrika*, Vol. IX) ; vide W. F Sheppard, loc. cit.

The fact that the computation of the mean square contingency leads to the same expression for the tetrachoric table has been stressed by K. Pearson simultaneously with the statement of the notion of the mean square contingency ; vide K. Pearson, *On the Theory of Contingency and its Relation to Association and Normal Correlation*, p. 21 (Drapers' Company Research Memoirs, Biometric Series, I ; 1904).

§ 2, 2. Various dodges of which one usually makes use at the statistical assessment of qualitative attributes are instructively reviewed in the final Chapters, XVI–XIX, of A. Niceforo's textbook, *Il Metodo Statistico. Teoria e Applicazione alle Scienze Naturali, alle Scienze Sociali, e all'Arte* (Messina, 1923).

§ 3, 1. The notion, ' Problème des Moments ' is due to Stieltjes, ' Recherches sur les Fractions Continues ', p. 48

(*Annales de la Faculté de Toulouse*, 1894). With regard to recent literature concerning the problem, G. Pólya, ' Über den Zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem ' (*Math. Zeitschrift*, Band VIII ; Berlin, 1920) and M. Riesz, ' Sur le Problème des Moments et le Theorème de Parseval Correspondent ' (*Skandinavisk Aktuarietidskrift*, 1924), should be mentioned. Further, vide J. F. Steffensen, *Matematisk Iagttagelseslaere*, pp. 41–2 (Köbenhavn, 1923).

§ 3, 1. The connexion $r_{1|1}^2 \leq 1$ can be derived in various ways. The proof mentioned in the text represents Yule's derivation in an *a priori* conception ; vide G. U. Yule, ' On the Significance of Bravais' Formulae for Regression . . . in the Case of Skew Correlation ', p. 482 (*Proc. Roy. Soc.*, Vol. 60 ; 1897). One can proceed likewise from $\mathsf{E}\left[\frac{x - m_{1|0}}{\sigma_x} \mp \frac{y - m_{0|1}}{\sigma_y}\right]^2 \geq 0$ (vide L. Tr. Kelley, *Statistical Methods*, p. 191) or from

$$\mu_{2|0}\mu_{0|2} - \mu_{1|1}^2 = \sum_i \sum_j [(x_i - m_{1|0})(y_j - m_{0|1}) - (x_j - m_{1|0})(y_i - m_{0|1})]^2$$

(vide A. L. Bowley, *Elements of Statistics*, p. 354). Further, the identity

$$\mu_{2|0}\mu_{0|2} - \mu_{1|1}^2 = \mu_{2|0} \sum_i p_{i|} \mu_{|2}^{(i)} +$$

$$+ \sum_i \sum_f \sum_h p_{i|}^2 p_{f|} p_{h|} \{[x_i - x_f][m_{|1}^{(i)} - m_{|1}^{(h)}] - [x_i - x_h][m_{|1}^{(i)} - m_{|1}^{(f)}]\}^2 + \sum_i \sum_g \sum_f \sum_h p_{i|} p_{g|} p_{|f} p_{|h} \{[x_i - x_g][m_{|1}^{(f)} - m_{|1}^{(h)}] - [x_f - x_h][m_{|1}^{(i)} - m_{|1}^{(g)}]\}^2$$

can be laid down as a basis (cf. my treatise in the second volume of the *Investigations by Russian Scientists Abroad*) which gives simultaneously an insight into the relationship between the correlation coefficients and correlation ratios. Finally, we may arrive at the proof that $r_{1|1}^2 < 1$ indirectly by $r_{1|1}^2 \leq \eta_{y|x}^2, \eta_{y|x}^2 \leq 1$ (cf. Chap. IV, § 4, 2 and § 4, 3). This shape of proof must be suggested particularly in the case of more than two correlated variables for the derivation of analogue relationship for the so-called coefficient of multiple correlation.

In a similar way it can be proved that the empirical correlation coefficient (cf. Chap. V, § 2, 2, § 4, § 5) in its absolute magnitude cannot be larger than 1.

§ 3, 2. (Cf. Chap. V, § 3.) Vide K. Pearson, *On the General Theory of Skew Correlation and Non-Linear Regression* (Drapers' Company Research Memoirs, Biometric Series, II ; 1905) ; vide K. Pearson, ' On a General Method of Determining the Successive Terms in a Skew Regression Line ' (*Biometrika*, Vol. XIII).

§ 3, 4. (Cf. Chap. V, § 3.) The presentation in the text corresponds, in *a priori* conception, to the methods of consideration due to Yule ; vide G. U. Yule, ' On the Significance of Bravais' Formulae for Regression . . . ', loc. cit. ; G. U. Yule, ' On the Theory of Correlation ', loc. cit.

§ 4, 3. Trials with partly remaining dice are eminently suitable for the illustrations in classes concerned with the importance of correlation coefficient as well as mutual relationship between the empirical correlation coefficients and their basic *a priori* correlation coefficients. A. D. Darbishire's ' Some Tables for Illustrating Statistical Correlation ' (*Mem. and Proc., Manchester Lit. and Phil. Soc.*, Vol. 51, No. 16, p. 1 ; 1907) may be regarded as a really rich source of such illustrations.

CHAPTER V

The limitation to the consideration to discontinuous distributions allows us to proceed in theoretical analysis from the assumption that the observed values of variables can be submitted to treatment without being contracted into classes. The problems which arise from the classification, which in the case of continuous distributions belong to the nature of the task of relevant shaping of empirical material, have, in the case of discontinuous distributions, only practical interest. For this reason I am not entering more closely into these problems, viz. into so-called Sheppard's corrections. With regard to this latter, I may refer to E. Pairman and K. Pearson, ' On Corrections for the Moment-Coefficients of Limited Range Frequency-distributions when there are finite or Infinite Ordinates and any Slopes at the Terminals of the Range ' (*Biometrika*, Vol. XII).

CHAPTER VI

§ 1 and § 2. Vide A. Tschuprow, ' Ziele und Wege der Stochastischen Grundlegung der Statistischen Theorie ', § 5, loc. cit.

§ 3, 2. (Cf. Chap. IV, § 2 ; Chap. V, § 7.) The magnitude δ' is called by K. Pearson ' the transfer per unit ' (K. Pearson, ' On the Correlation of Characters Not Quantitatively Measurable ', loc. cit., p. 14). For the theory of coefficients based on the magnitude δ, apart from Sheppard's and Pearson's referred to in considering Chapter I, § 2, 1, Yule's articles should be considered of first importance, from which we here mention only G. U. Yule, ' On the Association of Attributes in Statistics ' (*Phil. Trans.*, A, Vol. 194 ; 1900), and G. U. Yule, ' On the Methods of Measuring Association Between Two Attributes ' (*J. Roy. Stat. Soc.*, Vol. 75 ; 1912). Vide A. Tschuprow ' On Mathematical Expectation of Quotients of Two Correlated Chance Variables ', pp. 263 *et seq.* (Russian ; *Investigations by Russian Scientists Abroad*, Vol. I ; Berlin, 1922) ; V. Romanovsky, ' On Probabilities of Correlated Characteristics ' (Russian ; *Westnik Statistiki*, No. 12 ; Moscow, 1922).

§ 4. K. Pearson has dealt with the problem of calculation of $E\frac{x}{y}$ in his treatise, ' On a form of Spurious Correlation which may Arise when Indices are Used in the Measurement of Organs ' (*Proc. Roy. Soc.*, Vol. 60 ; 1897) ; vide K. Pearson, ' On the Constants of Index-Distributions as Deduced from the Like Constants for the Components of Ratio, with Special Reference to the Opsonic Index ' (*Biometrika*, Vol. VII). Later it was adapted by various investigators, who, in a similar way, but without having been aware of Pearson's treatise, derived some of Pearson's formulae ; vide E. Czuber, ' Über Funktionen von Variablen, Zwischen Welchen Korrelationen Bestehen ' (*Metron*, Vol. I ; 1920). A systematic survey of methods available at the time is described in my treatise, ' On the Mathematical Expectation of Quotients of Two Correlated Chance Variables ', referred to above.

§ 4, 1, A. Vide A. Tschuprow ' On the Mathematical Expectation of Quotients ', pp. 244–5. The form $E\frac{n_{i|j}}{n_{i|}} = \frac{p_{i|j}}{p_{i|}}$ following connexion $Em_{|1}^{(i)\prime} = m_{|1}^{(i)}$ has been approximately derived by K. Pearson in ' On the Application of " Goodness of Fit " Tables to Test Regression Curves and Theoretical Curves used to Describe Observational and Experimental Data ', p. 240 (*Biometrika*, Vol. XI) ; vide the editorial article ' Peccavimus ! ', p. 267

(*Biometrika*, Vol. XIII), where it is proved in a way different from mine that it holds good ' not merely to a high order of approximation, but absolutely '.

§ 4, 1, B. Vide A. Tschuprow ' On the Mathematical Expectation of Quotients ', pp. 246–51.

§ 4, 2. (Cf. Chap. IV, § 2, 1, and Chap. V, § 6.) Vide A. Tschuprow ' On the Mathematical Expectation of Quotients ', pp. 256–61. The coefficient φ (mean square contingency) is due to K. Pearson, ' On the Theory of Contingency and its Relation to Association and Normal Correlation ', loc. cit. The various formations of coefficients in the first instance are not differentiated sufficiently sharply. K. Pearson differentiates: ' the mean square contingency for the whole population ' (my φ), ' the approximate value of the mean square contingency ' (my φ), and ' its true value ', defined by the $\sum_i \sum_j \frac{[p'_{i|j} - p_{i|}p_{|j}]^2}{p_{i|}p_{|j}}$. Vide K. Pearson ' On the Probable Error of a Coefficient of Mean Square Contingency ' (*Biometrika*, Vol. X), and K. Pearson and A. W. Young, ' On the Probable Error of a Coefficient of Contingency Without Approximation ' (*Biometrika*, Vol. XI; some of the formulae of this treatise are not quite correct; vide ' Peccavimus ! ', pp. 259–60). The probable error of φ' is to the first approximation derived in a different way from mine by J. Blakeman and K. Pearson, ' On the Probable Error of Mean Square Contingency ' (*Biometrika*, Vol. V).

§ 4, 3, A. Vide A. Tschuprow, ' On the Mathematical Expectation of Quotients ', pp. 267–9. For the case of normal correlation systematic errors and the probable error of empirical correlation coefficient is derived in a similar way by H. E. Soper, ' On the Probable Error of a Correlation Coefficient to a Second Approximation ' (*Biometrika*, Vol. IX). R. A. Fisher, inspired by the treatise by Student's, ' Probable Error of a Correlation Coefficient ' (*Biometrika*, Vol. VI), has derived the law of distribution of values of empirical coefficients of correlation for the case of normal correlation; R. A. Fisher, ' On the Probable Error of a Coefficient of Correlation deduced from a Small Sample ' (*Metron*, Vol. I; 1921). The general formula for the probable error of empirical coefficient is due to W. F. Sheppard, loc. cit., p. 128; the formula mostly used for the probable error of $r_{1|1}$

in the case of normal correlation takes its original from K. Pearson and L. N. G. Filon, ' On the Probable Errors of Frequency-Constants and on the Influence of Random Selection on Variation and Correlation ', p. 245 (*Phil. Trans*, A, Vol. 191 ; 1898). In older literature we occasionally find the value of standard error of empirical correlation coefficient denoted by $\frac{1 - r_{1|1}^2}{\sqrt{N(1 + r_{1|1}^2)}}$. This formula is due to K. Pearson, ' Regression, Heredity, and Panmixia ', loc. cit., p. 266 ; and is replaced in Pearson and Filon's above-mentioned treatise by the correct approximation formula.

§ 4, 3, B. The probable error of $A'_{|1}$ is, under assumption of normal correlation, derived by Pearson and Filon (loc. cit., p. 245) ; in the editorial treatise, ' On the Probable Errors of the Frequency-Constant ', p. 9 (*Biometrika*, Vol. IX), the formula, after dropping the assumption of normal correlation, is reproduced under the supposition that ' the frequencies are symmetrical and the regression linear ' holds good ; in the same place the standard error of $A'_{|0}$ is stated in the form

$$\delta_{A'_{|1}} = \sqrt{[m_{1|0}^2 + \mu_{2|0}]}.$$

§ 4, 3, C. (Cf. Chap. IV, § 4, and Chap. V, § 4.) Vide V. Romanovsky, ' On Correlation Ratio ' (Russian ; *Westnik Statistiki*, No. 12 ; Moscow, 1922). The Magnitude η' is introduced by K. Pearson, *On the General Theory of Skew Correlation and Non-Linear Regression* (Drapers' Company Research Memoirs, Biometric Series, II ; 1905). In this treatise the probable error of η' is stated on page 19. The systematic estimation error of η' was calculated at a later date ; cf. K. Pearson, ' On the Correction Necessary for the Correlation Ratio η ' (*Biometrika*, Vol. XIV).

§ 4, 3, D. Vide J. Blakeman, ' On Tests for Linearity of Regression in Frequency-Distributions ' (*Biometrika*, Vol. IV).

§ 4, 5. The probable error of $\frac{\delta'}{\sqrt{p'_{1|}p'_{2|}p'_{|1}p'_{|2}}}$ is calculated by G. Yule, ' On the Methods of Measuring Associations Between Two Attributes ', loc. cit., p. 603, in a manner deviating from mine.

§ 5. Vide E. Slutsky, ' On Some Schemes of Correlated Con-

nexion and Systematic Error of Empirical Correlation Coefficient ' (Russian ; *Westnik Statistiki*, No. 13 ; Moscow, 1923). Vide the presentation of the problem by L. Isserlis, ' The Variation of the Multiple Correlation Coefficient in Samples, Drawn from an infinite Population with Normal Distribution ' (*Phil. Mag.*, 6th Series, Vol. 34 ; 1917).

CHAPTER VII

The conception stated here is developed in greater detail in my treatise, ' Ziele und Wege der Stochastischen Grundlegung der Statistischen Theorie ', §§ 7, 8 ; ' Über Normal-stabile Korrelation ' (loc. cit.). Vide J. Morduch, ' On Connected Trials which Meet the Condition of Stochastic Commutability ' (Russian ; *Investigations by Russian Scientists Abroad*, Vol. II ; Berlin, 1923).

CHAPTER VIII

§ 1, 2. With regard to the application of coefficients it is very instructive to compare the treatise by G. U. Yule, ' On the Association of Attributes in Statistics ' (*Phil. Trans.*, A, Vol. 194 ; 1900), and ' On the Methods of Measuring Association Between Two Attributes ', loc. cit., with the treatise by K. Pearson and D. Heron referred to above.

§ 2, 1. As examples of cartographical consideration of correlation measurements in the meteorology may be mentioned F. Exner, ' Über Monatliche Witterungsanomalien auf der Nördlichen Erdhälfte im Winter ' (*Sitz.-Ber. d. Akademie d. Wissensch.*, Vol. 122, Sekt. IIa ; Vienna, 1913 ; reviewed with reproduction of illustrations by A. Defant, *Wetter und Wettervorhersage* (Leipsic and Vienna, 1918), and by A. Schmaus, ' Korrelationen von März bis September ' (*Meteor. Zeitschrift*, Vol. 41 ; 1924).

Vide further, ' Brit.-Antarctic Exped., 1910–13 ', *Meteorology*, Vol. I (Calcutta, 1919).